PANACM 2015
1st Pan-American Congress on Computational Mechanics

In conjunction with the XI Argentine Congress on Computational Mechanics, MECOM 2015

27-29 April, 2015, Buenos Aires, Argentina
1st Pan-American Congress on Computational Mechanics - PANACM 2015

XI Argentine Congress on Computational Mechanics - MECOM 2015
Proceedings of the

1st Pan-American Congress on Computational Mechanics

and

XI Argentine Congress on Computational Mechanics

held in Buenos Aires, Argentina
27-29 April 2015

Edited by:

Sergio R. Idelsohn (Chairman)
International Centre for Numerical Methods in Engineering (CIMNE)
- Catalan Institute for Research and Advanced Studies (ICREA),
  Barcelona, Spain
Centro de Investigaciones de Métodos Computacionales (CIMEC),
Santa Fe, Argentina

Victorio Sonzogni
Centro de Investigaciones de Métodos Computacionales (CIMEC),
Santa Fe, Argentina

Alvaro Coutinho
High Performance Computing Center (COPPE)
Federal University of Rio de Janeiro, Brazil

Marcela Cruchaga
Universidad de Santiago de Chile, Chile

Adrian Lew
Stanford University, Stanford (California), USA

Miguel Cerrolaza
Instituto Nacional de Bioingeniería de la Universidad Central de Venezuela
Caracas, Venezuela

A publication of:

International Center for Numerical Methods in Engineering (CIMNE)
Barcelona, Spain
1st Pan-American Congress on Computational Mechanics

and

XI Argentine Congress on Computational Mechanics

First edition, April 2015

© The authors

Printed by: Artes Gráficas Torres S.L., Huelva 9, 08940 Cornellà de Llobregat, Spain

E-book:
PREFACE


PANACM 2015 is the first conference organized under the auspices of the International Association for Computational Mechanics (IACM) to promote the Computational Mechanics in the Americas as a one region.

The purpose of PANACM series is to promote achievements by encouraging young researchers, stimulating education in universities, disseminating modern trends in the field amongst scientists and engineering and mainly, facilitating the interchange of knowledge between the North and the South.

The main objective of PANACM 2015 is to become a forum for state of the art presentations and discussions of mathematical models, numerical methods and computational techniques for solving problems of multidisciplinary character in science and engineering. The conference goal is to make a step forward in the formulation and solution of real life problems with a multidisciplinary nature and industrial interest, accounting for all the complex mathematical models involved in the physical description of the problem.

PANACM 2015 is organized in conjunction with the XI Argentine Congress on Computational Mechanics, MECOM 2015, a traditional and very well known congress in the Argentinean science community.

PANACM 2015, together with MECOM 2015, has attracted around 600 participants, coming from all over the world. All together some 540 lectures will be presented, including 8 Plenary lectures, 12 Semi-Plenary lectures and 43 Keynote Lecturers, which reflect the current state of the research and advances in engineering practice in the Computational Mechanics science.

The International Centre for Numerical Methods in Engineering (CIMNE) organizes this Conference jointly with the Asociación Argentina de Mecánica Computacional (AMCA). The organizers acknowledge the encouragement and support of CIMNE and AMCA and the special interest of the IACM, under whose auspices this conference is held.


The organizers would like to thank the authors for submitting their contributions and for their respect of the deadlines. Special thanks go to the colleagues who contributed to the organization of the 45 Mini-Symposia to be presented during the Congress.

Sergio Idelsohn (Chairman)
Victorio Sonzogni (Argentina)
Alvaro Coutinho (Brazil)
Marcela Cruchaga (Chile)
Adrian Lew (USA)
Miguel Cerrolaza (Venezuela)
ACKNOWLEDGEMENTS

PANACM 2015 is supported by the following IACM American Affiliated Associations:

**Affiliated Associations**

- AMCA - Asociación Argentina de Mecánica Computacional.
- Asociación Mexicana de Modelación Numérica e Ingeniería (AMMNI)
- Brazilian Association for Computational Methods in Engineering (ABMEC)
- Colombian Association for Numerical Methods in Engineering and Applied Sciences (ACMENICA)
- Sociedad Chilena de Mecánica Computacional (SCMC)
- U.S. Association for Computational Mechanics (USACM)
- Venezuelan Society for Numerical Methods in Engineering (SVMNI)
Supporting Organizations

The congress organizers acknowledge the support towards the organization of the PANACM 2015 Congress to the following organizations:

- Agencia Nacional de Promocion Científica y Tecnologica de Argentina
- ASAMACI - Asociación Argentina de Matemática Aplicada Computacional e Industrial
- Buenos Aires Convention Bureau
- CIMEC- Centro de Investigaciones de Métodos Computacionales.
- CIMNE - International Center for Numerical Methods in Engineering
- CIMNE - Latin American Foundation (FCL)
- CONICET - Consejo Nacional de Investigaciones Científicas y Técnicas
- IACM - International Association of Computational Mechanics
- ICREA - Institució Catalana de Recerca i Estudis Avançats
- Universidad Nacional del Litoral
PLENARY AND SEMI-PLENARY SPEAKERS

Plenary Speakers: Klaus Jürgen Bathe, Luis Caffarelli, Charbel Farhat, Thomas J. R. Hughes, Eugenio Oñate, Olivier Pironneau, Ekkehard Ramm and Peter Wriggers.

Semi-Plenary Speakers: Gustavo Buscaglia, Alberto Cardona, Leszek Demkowicz, Eduardo N. Dvorkin, Irene M. Gamba, Antonio Huerta, Abimael Loula, Xavier Oliver, Rodolfo Rodriguez, Mario Storti, Frédéric Valentin and Tarek Zohdi.

MINI-SYMPOSIA ORGANIZERS

1.3 MS: Advances in Numerical Methods for Multibody Systems
   Organized by Alberto Cardona, Olivier Brüls and Martin Pucheta

1.8 MS: Computational Buckling Analysis of Thin-Walled Structures
   Organized by Luis Godoy and Rossana Jaca

1.13 MS: Computational Geomechanics
   Organized by Dorival Pedroso, Jose E. Andrade and Patricia Faria

1.18 MS: Computational Methods in Seismic Engineering
   Organized by Miguel Eduardo Tornello, Oscar Möller and Francisco J. Crisafulli

1.23 MS: Large Deformation Problems. Fundamentals and Applications
   Organized by Carlos Garcia Garino, Jean-Philippe Ponthot and Anibal E. Mirasso

1.25 MS: Material Modeling with Multiscaling
   Organized by Kenjiro Terada, Eduardo de Souza Neto, Pablo Blanco and Peter Wriggers

1.28 MS: Multiscale Computational Design of Engineering Materials
   Organized by Alfredo Edmundo Huespe, Pablo Sánchez, Guillermo Etse and Xavier Oliver

1.29 MS: Nonlocal Models for Mechanics and Diffusion
   Organized by Pablo Seleson, Diego Del-Castillo-Negrete and Michael Parks

1.30 MS: Numerical Modeling of Damage Evolution and Propagating Discontinuities in Solids at Failure
   Organized by Christian Linder and Ekkehard Ramm

1.32 MS: Structural Analysis and Vibrations
   Organized by Diana V. Bambill and Liz Nallim

1.33 MS: Structural Integrity – Fracture Mechanics
   Organized by Hiroshi Kawai and Hiroshi Okada

1.34 MS: Structures under Extreme Loads
   Organized by Daniel Ambrosini and Bibiana Luccioni
2.5 MS: Computational Fluid Dynamics and Finite Element Methods: Formulations, Implementation Strategies and Applications

Organized by Lucia Catabriga and Sandra Malta

2.8 MS: Computer Methods for Environmental Fluid Mechanics

Organized by Angel N. Menéndez, Renato Elias and Kazuo Kashiyama

2.12 MS: Coupled Problems and Free Surface Flows

Organized by Facundo Del Pin and Rodrigo Paz

2.15 MS: DNS/LES of Turbulent Wall-Bounded Flows

Organized by Xiaohua Wu and Guillermo Araya

2.19 MS: Free-Surface, Moving-Boundary and Multi-Phase Flows

Organized by Marcela Cruchaga, Mario Storti, Gustavo Buscaglia and Axel Larrenguy

2.27 MS: Numerical Modeling of Flooding Areas and Erodible Bathymetry

Organized by Pablo Ortiz and Joanna Szmelter

3.1 MS: A Priori and A Posteriori Error Analysis in Finite Element Methods

Organized by Ricardo Duran

3.2 MS: Advanced Numerical Strategies for the Design of Materials, Structures and Processes

Organized by Francisco Chinesta, Elias Cueto and Antonio Huerta

3.3 MS: Algorithmic Aspects of High-Performance Computing for Mechanics and Physics

Organized by Santiago Badia, Victor Calo and Javier Principe

3.5 MS: Boundary Element Techniques

Organized by Euclides Mesquita, Luis Tavara and Adrián Pablo Cisilino

3.7 MS: Computational Methods for Kinetic Collisional Transport

Organized by Irene M. Gamba and Thierry Magin

3.12 MS: Finite Element Methods for Multiscale Problems

Organized by Frédéric Valentin and Alexandre Madureira

3.14 MS: Free Boundary Problems: Theory, Numerical Analysis and Applications

Organized by Domingo Tarzia

3.17 MS: High Performance Computing and Related Topics

Organized by Ryuji Shioya and Daisuke Tagami

3.19 MS: Higher Order Finite Element Discretizations

Organized by Leszek Demkowicz, Jay Gopalakrishnan and Tan Bui-Thanh

3.20 MS: HPC Techniques in Computational Mechanics

Organized by Mariano Vázquez, Alvaro Coutinho and Mario Storti

3.25 MS: Mathematical and Computational Methods for Subsurface Modeling and Visualization

Organized by Víctor Calo, Ignacio Muga and David Pardo

3.26 MS: Mathematics and Algorithms for Advanced Numerical Simulations

Organized by Rodolfo Araya, Carlos Conca, Pascal Frey and Olivier Pironneau
3.30 MS: Model Order Reduction of Large-Scale Mechanical Systems
   Organized by David Amsallem and Charbel Farhat

3.33 MS: Numerical Analysis of Partial Differential Equations in Continuum Mechanics and Electromagnetism
   Organized by Rodolfo Rodriguez

3.37 MS: Recent Advances in Meshfree and Particle Methods
   Organized by Seiichi Koshizuka, Kazuo Nishimoto and Liang-Yee Cheng

4.1 MS: Advances in Computational Biomechanics
   Organized by Daniel Hurtado and Reuben Kraft

4.3 MS: Computational Mechanics in Biology and Medicine
   Organized by Sandra Rugonyi and Suvranu De

4.8 MS: Numerical and Computational Methods in Analysis and Design of Biomedical Devices
   Organized by José Di Paolo, Marcelo E. Berli and Miguel Cerrolaza

5.6 MS: Inverse Problems and their Applications
   Organized by Ruben D. Spies

5.12 MS: Optimization and Computational Mechanics
   Organized by Alfredo Canelas, Jose Herskovits Norman, Sandro Mazorche and Mario Tanaka Filho

5.13 MS: Robust, Performance-based and Reliability-based Optimization under Uncertainty
   Organized by André Beck, Marcos Valdebenito and Rafael Holdorf Lopez

5.16 MS: Uncertainty Modeling and Quantification
   Organized by Rubens Sampaio, Marta Rosales and Marcelo Tulio Piovan

5.17 MS: Uncertainty Quantification and Inverse Problems
   Organized by Hermann G. Matthies, Youssef Marzouk, Habib Najm, Alireza Doostan and Sondipon Adhikari

6.5 MS: High Order Methods with Applications to Aerodynamics and Aeroacoustics
   Organized by Joao Luiz Azevedo and William Wolf

8.2 MS: Aerospace Technology
   Organized by Sergio Elaskar and Eduardo Zapico

8.5 MS: CFD in Engineering and Technology of Nuclear Installations
   Organized by Juan Carlos Ferreri and Alejandro Iván Lazarte

8.10 MS: Industrial Applications of Computational Mechanics
   Organized by Marcela Goldschmit and Jose Risso
CONTRIBUTIONS INDEX

1 - Computational Solid Mechanics

1.1 Computational Buckling Analysis of Thin-Walled Structures

Keynote Lecturer: Luis Godoy
A penalty approach to obtain lower bound buckling loads of imperfection-sensitive shells .. 29
L. A. Godoy*, R. C. Jaca, E. M. Sosa and F. G. Flores

Buckling of composite cross-ply sandwich panels using different finite element .......... 41
4141approximations
F. G. Flores*

1.2 Computational Geomechanics

Interaction between soil and buried pipeline ............................................................... 53
S. M. F. S. Melo, L. M. da Costa* and I. Pontes

Finite element analysis of mechanical deformation processes in sedimentary basins ....... 63
A. Brüch* and S. Maghous

A non-isothermal consolidation model for gradient-based poroplasticity ....................... 75
J. L. Mroginski*, G. Etse and M. Ripani

Modelling fluidisation and sedimentation using material point method ....................... 89
M. Martinelli* and A. Rohe

Recent development in the Combined Finite-Discrete Element Method ....................... 101
E. Rougier*, E. E. Knight, Z. Lei, A. Munjiza

1.3 Computational Methods in Seismic Engineering

Structural behavior of unreinforced masonry piles of different thicknesses subject to
compression loads ........................................................................................................ 112
N. D. Agüera*, M. E. Tornello and C. D. Frau

Influence of the parameters characterizing the near-fault earthquake in the
seismic response of base isolation structures ................................................................ 122
M. E. Tornello*, C. D. Frau and D. Pagano

Performance-based design optimization of structures under seismic demands,
incorporating passive control devices ........................................................................ 132
O. Möller*, R. O. Foschi and J. P. Ascheri

Concepts of dampers for earthquake protection of existing buildings
and for displacements restraints in seismically isolated buildings .............................. 144
M. G. Melkumyan*

1.4 Large Deformation Problems. Fundamentals and Applications

Atomistic study of the mechanical properties of a sintered bulk metallic glass (Nanoglass) ....156
F. Ardiani*, A. A. Manelli, C. J. Ruestes, C. A. Careglio and E. M. Bringa
Mechanical properties of a Cu$_{46}$Zr$_{54}$ bulk metallic glass with embedded crystalline nano particles ................................................................. 166
A. A. Manelli*, F. Ardiani, C. A. Careglio and E. M. Bringa

Numerical simulation of elastic springback in problems including large elastic-plastic strains ................................................................. 177
W. B. Castelló* and F. G. Flores

1.5 Material Modeling with Multiscaling

A FFT based mesoscopic approach for the compression and recovery simulation of structured nonwovens ................................................................. 189
S. Staub*, H. Andrä and M. Kabel

Validation of multiscale model for heat generation in hardening concrete...................... 199
W. R. Leal da Silva, V. Šmilauer and Z. Bittnar*

Computational material modeling of hydrated cement paste calcium silicate hydrate (C-S-H) chemistry structure - influence of magnesium ion exchange on mechanical stiffness: C-S-H Jennite ................................................................. 205
R. V. Mohan*, W. D. Hodo, B. Adebiyi, J. Rivas and A. Mohamed

A multi-scale electro-chemo-mechanical analysis of solid oxide fuel cell considering evolution of microstructure in porous electrode ............................................ 215
M. Muramatsu*, S. Takase, K. Yashiro, T. Kawada and K. Terada

1.6 Multiscale Computational Design of Engineering Materials

Localized versus diffused failure modes in concrete subjected to high temperature ...... 225
S. M. Vrech, M. Ripani* and G. Etse

Structural efficiency of multiscale topology optimization of cellular materials using Bi-directional Evolutionary Structural Optimization......................................................... 237
T. K. L. Calixto* and R. Pavanello

Continuous microplane theory and interface approach for failure form analysis of steel fiber reinforced concrete ............................................................................. 249
S. M. Vrech*, A. Caggiano and G. Etse

Proposal of an interpolation function between the compressive and tensile meridians of failure and yielding concrete surfaces based on Bezier curves ................... 261
P. Folino* and D. Smilovich

1.7 Advances in Numerical Methods for Multibody Systems

Analysis of the oscillation of a self-excited pendulum due to varying length ............... 272
M. C. Pereira* and H. I. Weber

Simulation of an active suspension using PID control .................................................. 281
D. F. Sousa* and S. M. Avila

Synthesis of multi-fingered robotic hands ..................................................................... 292
A. Perez Gracia*
1.8 Numerical Modeling of Damage Evolution and Propagating Discontinuities in Solids at Failure

Blade cutting simulation with crack propagation through thin-walled structures via solid-shell finite elements in explicit dynamics ................................................................. 304
A. Ghisi*, F. Confalonieri and U. Perego

Stability and error estimate of a cohesive zone model implemented using the augmented lagrangian method ................................................................. 314
N. A. Labanda*, S. M. Giusti and B. M. Luccioni

Multi-region Boundary Element Method and Tangent Operator Technique applied to crack propagation modelling in concrete structures ........................................... 327
S. G. F. Cordeiro* and E. D. Leonel

1.9 Structural Analysis and Vibrations

Keynote Lecturer: Diana V. Bambill

Free vibrations of Timoshenko micro beams using the modified couple stress theory ..... 340
G. I. Guerrero, D. V. Bambill* and D. H. Felix

Analysis of the behavior of an industrial building with presence of vibrations due to machinery ........................................................................................................ 352
N. L. Ercoli, M. H. Peralta* and L. O. Pico

A simplified numerical model for assessing the residual strength and deformation capacity of asymmetrically damaged ropes ........................................ 364
J. F. Beltrán* and E. De Vico

Comparative analysis of new shear locking-free finite element with other commonly used approaches in vibration analysis of Mindlin plates ....................... 375
I. Senjanovic, N. Vladimir*, D. S. Cho, M. Jokic and T. M. Choi

Dynamic analysis by a new family of time-marching procedures with adaptive time integration parameters ................................................................. 389
D. Soares*

Direct assessment of interlaminar stresses in composite multilayered plates using a layer-wise mixed finite element model .................................................. 398
R. S. Carnicer *, B. A. Desía and R. A. Schwarz

Free vibrations of an axially functionally graded beam with a concentrated mass ........ 410
G. J. Gilardi, D. V. Bambill and C. A. Rossit*

Towards a new joint element which incorporates bolt looseness and nonlinear contact behaviour .................................................................................................. 422
F. del Solar Dorrego* and S. D’hers

On the nonlinear dynamics of network structures under stochastic excitation .......... 434
W. de Cerqueira Leite* and R. M. L. R. F. Brasil

Size-dependent dynamic analysis of Timoshenko microbeam using the finite element method ......................................................................................... 446
D. H. Felix*, G. I. Guerrero and D. V. Bambill

A stress analysis app .................................................................................................. 458
S. Gallegos Câzares* and A. Balderrama López
Use of computational fluid dynamic to compare the pressure loss between a parallel flow field plate and a parallel-baffle flow field plate in a direct ethanol proton exchange membrane fuel cell ................................. 468
J. Vicente*, G. Benetti, M. Forte, G. Heusy, E. Mathias and P. Belchor

1.10 Structural Integrity – Fracture Mechanics

Hydraulic fracturing simulations using a homogenized continuum approach .............. 474
P. H. E. Cheroto*

Computational modeling of vehicle-irregularity-bridge dynamic interaction by damage mechanics .................................................................................................................. 485

1.11 Structures under Extreme Loads

Keynote Lecturer: Norbert Gebbeken
Protective Structures - A challenge to computational engineering and testing ............. 497
N. F. J. Gebbeken* and T. Bermbach

The analysis of splice joint behaviour in heavily loaded tensile chords of warren trusses ..... 507
L. Ozola*

Keynote Lecturer: Hong Hao
Finite element analysis of structural insulated panel with OSB skins against windborne debris impacts ........................................................................................................... 518
H. Hao*, W. Chen, S. Chen and Q. Meng

Ultimate loading capacity of RC columns under fire: Comparison of ACI and Eurocode Models ................................................................................................................ 525
H. Barros*, C. Ferreira and V. D. Silva

1.12 Advanced Methods in Computational Solid Mechanics

Vibration analysis and human comfort assessment of steel-concrete composite floors subjected to human rhythmic activities ................................................................. 531
C. M. R. Gaspar and J. G. Santos da Silva*

Shear-wall mock-up subjected to monotonic loading with concrete-steel bond model ..... 541
C. Mang*, L. Jason and L. Davenne

Zero-thickness interface formulation for fracture analysis of self-healing concrete ....... 553
A. Caggiano*, V. Krelani, L. Ferrara and G. Etse

1.13 Geometrical and Material Non-linear Problems

Approach to enhance computational robustness of frictional contact analysis .............. 565
M. R. Gurvich*

A new form of finite element method for plane stress/plane strain problems in reinforced concrete considering material nonlinearity and cracking .............................. 575
K. G. Ashkinadze*
2 - Computational Fluid Dynamics

2.1 Computational Fluid Dynamics and Finite Element Methods: Formulations, Implementation Strategies and Applications

CFD applications on the design and analysis of aeronautical products at FAdeA SA ........ 587
G. Mognol and D. M. Carrasco*

Comparison between the drag polar curves using analytical and numerical methods in controlled radio airplane ................................................................. 599
A. P. Almeida, L. M. Pereira, C. C. G. Santos and L. S. Machado*

2.2 Computer Methods for Environmental Fluid Mechanics

Parallel computations for many floating objects transported by tsunami flows ............ 611
K. Aoki, S. Ushijima, H. Itada* and D. Toriu

Practical criteria for numerical modeling of water quality in coastal zones ............... 623
A. N. Menéndez*, N. D. Badano, E. A. Lecertúa and M. Re

Numerical modeling of heterogeneous media for the compressional wave using High-Order Finite-Differences .......................................................... 635
A. Guillen, F. H. Cabrera, J. A. Torres and J. E. Duran*

2.3 Coupled Problems and Free Surface Flows

Computational method for interactions between compressible fluids and solids with thermal conductivity ................................................................. 647
D. Toriu* and S. Ushijima

2.4 Free-Surface, Moving-Boundary and Multi-Phase Flows

Numerical simulation of the Weissenberg effect......................................................... 658
R. A. Figueiredo, C. M. Oishi and J. A. Cuminato*

XFEM modelling of hydraulic fracturing ................................................................ 669
J. B. Lucero, M. Luege, A. Orlando* and C. Torrijos

Global Method of Approximate Particular Solutions for solving the rectangular two-sided lid-driven cavity problem for two-dimensional and incompressible flow ........ 681
J. M. Granados*, C. A. Bustamante, H. Power, W. F. Florez and C. Nieto

2.5 Numerical Modeling of Flooding Areas and Erodible Bathymetry

Model reduction for shallow water flows ................................................................. 693
M. W. Farthing*, C. E. Kees and A. Lozovskiy

2.6 Advanced Methods in Computational Fluid Dynamics

Ventilation system for drying out walls after a flood: influence of the composition of the walls ................................................................. 705
I. Torres* and Ó. Lopez
Modelling the air-fuel mixture in an atmospheric burner ............................................. 717
A. C. Sá, S. F. Teixeira and J. C. Teixeira*

Computational fluid dynamic analysis of olive oil in different induction pots ............. 729
J. P. Kastillo*, J. Martinez, A. J. Riofrío, S. Villacis and M. Orozco

Numerical studies of Paraguayan Chaco natural gas in a counterflow diffusion flame ...... 742
H. A. Maldonado*, D. A. Alviso, M. A. Mendieta and J. C. Rolon

Shape optimization of an axisymmetric blunt body in hypersonic flow for reducing
drag and heat transfer ................................................................. 754
R. K. Agarwal* and C. Seager

A novel algebraic extra-stress model and its analytic solution for fully
developed laminar flows ................................................................ 767
G. S. Paulo*, G. Mompean, C. M. Oishi and M. F. Tomé

Solution of two-dimensional Shallow Water Equations by a localized Radial
Basis Function collocation method ...................................................... 775
C. A. Bustamante*, H. Power, C. Nieto and W. F. Florez

3 - Computational Mathematics and Numerical Methods

3.1 A Priori and A Posteriori Error Analysis in Finite Element Methods

Estimates and corrector based mesh adaptation ..................................................... 787
G. Brèthes*, A. Loseille, F. Alauzet and A. Dervieux

3.2 Advanced Numerical Strategies for the Design of Materials, Structures
and Processes

A Secant-Stepback procedure for the solution of severe geometric
and physical nonlinear problems .................................................................. 799
M. H. A. Lima Jr.*, F. N. Corrêa, B. P. Jacob, B. M. Jacovazzo and B. F. Monteiro

3.3 Algorithmic Aspects of High-Performance Computing for Mechanics and
Physics

A variational flux recovery approach for elastodynamics problems with interfaces ....... 811
P. Bochev* and P. Kuberry

3.4 Boundary Element Techniques

Level set analysis of topology optimization in 2D structures using Boundary
Element Method .................................................................................... 823
P. C. Vitorio Junior* and E. D. Leonel

3.5 Finite Element Methods for Multiscale Problems

A new multiscale simulation of water flows in heterogeneous porous media ............... 834
A. Francisco*, J. Trujillo and L. Sacramento
3.6 Free Boundary Problems: Theory, Numerical Analysis and Applications

Keynote Lecturer: Domingo Tarzia
Neumann solutions to fractional Lamé-Clapeyron-Stefan problems with heat flux or convective boundary conditions ................................................................. 849
D. A. Tarzia*

Numerical solutions for free boundary problems with either shrinkage or expansion ........ 859
M. C. Olguin*, E. A. Santillan Marcus and M. C. Sanziel

Determination of two unknown thermal coefficients through a mushy zone with a convective overspecified boundary condition ........................................... 869
A. N. Ceretani* and D. A. Tarzia

Cumulative nutrient uptake by roots of crops as simulated by fixed and moving boundary models. Corrections and improvements .............................. 876
J. C. Reginato*, J. L. Blengino and D. A. Tarzia

3.7 High Performance Computing and Related Topics

Study of effects of blood amino acid and hormone level for controlling triglyceride accumulation in the liver of rats using Self-Organizing Map ...................................................... 885
M. Masuda*, Y. Nakabayashi, R. Shioya, H. Nishi, S. Takahashi and F. Hakuno

Accelerated multi GPU tsunami run-up simulation .............................................. 892
S. Furuyama*, Y. Maihara and T. Aso

Varying the block size on block conjugate gradient: comparison of strategies .......... 902
P. J. Torres*, C. E. Schaerer and A. Bhaya

A domain-decomposition based parallel procedure for the combined finite-discrete element method in 2D ................................................................. 912
E. E. Knight, Z. Lei, E. Rougier*, A. Munjiza

3.8 Higher Order Finite Element Discretizations

Application of discontinuous Galerkin adaptive mesh and order refinement method to energy transport and conservation equation in radiation-hydrodynamics .......................... 919
M. Holec*, M. Cotelo, P. Velarde and R. Liska

Multi-frontal multi-thread direct solver with GALOIS system for adaptive finite element method ................................................................. 931

Use of maximum entropy shape function within a strong meshfree formulation .......... 943
F. A. Marchant* and F. O. Perazzo

3.9 HPC Techniques in Computational Mechanics

A trade-off analysis between high-order seismic RTM and computational performance tuning ....955
A new parallel numerical model for multiple collisions .................................................. 963
R. Leon*, L. Salinas, C. Torres and G. Hernandez

Parallel Delaunay tetrahedrization of a given discrete point set constrained by an imposed boundary mesh ................................................................. 972
P. J. Novara* and N. A. Calvo

Assessment of Edge-based Finite Element technique for geophysical electromagnetic problems: efficiency, accuracy and reliability ........................................... 984
O. Castillo Reyes*, J. de la Puente, V. Puzyrev and J. M. Cela

Construction of school timetable with emphasis in minimization of twinned lessons and windows ................................................................. 996
M. I. S. Mikuska, C. E. Galvão Rosa *, A. R. T. Góes and S. Scheer

3.10 Mathematical and Computational Methods for Subsurface Modeling and Visualization

Keynote Lecturer: Helene Barucq

Performance assessments of absorbing conditions for the reverse time-harmonic migration ................................................................. 1007
H. Barucq*, M. Bergot, J. Chabassier and J. Diaz

3.11 Mathematics and Algorithms for Advanced Numerical Simulations

Droplet in micro-channels: a numerical approach using an adaptive two phase solver .. 1015
J-M. Fullana*, Y. Ling, S. Popinet and Ch. Josserand

3.12 Model Order Reduction of Large-Scale Mechanical Systems

Global reduced-order model adapted to the low- and medium-frequency analysis of complex dynamical structures ................................................................. 1022
O. Ezvan*, A. Batou and C. Soize

Essential boundary conditions and k-compressibility in Hyper Reduced Order Models .... 1029
A. Cosimo*, A. Cardona and S. R. Idelsohn

3.13 Numerical Analysis of Partial Differential Equations in Continuum Mechanics and Electromagnetism

Numerical simulation of ultrasonic wave propagation using higher order methods in space and time ................................................................. 1041
J. F. Unger*

3.14 Recent Advances in Meshfree and Particle Methods

The influence of geometrical parameters on the performance of perforated breakwater ...... 1053
R. Augusto Amaro Junior*, L-Y. Cheng and K. Nishimoto

Numerical simulation of vortex interactions using a fast multipole discrete particle method ................................................................. 1065
T. R. Ricciardi*, A. M. Bimbato and W. R. Wolf
3.15 Advanced Numerical Methods
An application of generalized differences to unsteady heat problems subject to mixed boundary conditions ................................................................. 1077
F. J. Domínguez-Mota*, J. G. Tinoco-Ruiz, F. O. Guillén-Reyes, G. Tinoco-Guerrero and A. Valencia-Ramírez

4 - Computational Methods in Life Sciences

4.1 Advances in Computational Biomechanics
Combined finite element and musculoskeletal models for analysis of pelvis throughout the gait cycle ................................................................. 1083
E. P. Ravera*, M. J. Crespo, F. A. Guarnieri and A. A. Braidot

Mechanical properties of foot tendons – In vitro study ........................................ 1095
E. Morales-Orcajo*, R. Becerro-de-Bengoa-Vallejo, M. Losa-Iglesias and J. Bayod

Using 1D-IRBFN method for solving high-order nonlinear differential equations arising in models of active-dissipative systems ........................................ 1103
D. V. Strunin*, D. Ngo-Cong and R. P. Bhanot

4.2 Computational Mechanics in Biology and Medicine
Numerical simulation of cooling and freezing processes for the cryopreservation of biological material in liquid nitrogen .................................................. 1111
M. V. Santos*, M. Sansinena, J. Chirife and N. Zaritzky

4.3 Numerical and Computational Methods in Analysis and Design of Biomedical Devices
Complex variable sensitivity analysis of the radio frequency ablation process for cancer treatment................................................................. 1124
J. F. Monsalvo S.* and M. J. García R.

Lattice-Boltzman modelling of porous medium: application to modelling Flow Diverter stents ........................................................................ 1133
G. A. Boroni*, J. A. Dottori, A. Clausse and I. Larrabide

5 - Optimization and Control

5.1 Inverse Problems and their Applications
Keynote Lecturer: Ruben D. Spies
Existence and stability of minimizers of mixed anisotropic BV-L² Tikhonov-Phillips functionals: applications to image restoration .................................................. 1144
G. L. Mazzieri, R. D. Spies* and K. G. Temperini

A method for atoms selection applied to screening for sleep disorders .................. 1155
Ad-hoc Gaussian dictionaries for sparse representation of Evoked Related Potentials .... 1167
V. Peterson*, H. L. Rufiner and R. D. Spies

Local edge-preserving image inpainting with mixed weighted anisotropic regularization methods ................................................................. 1179
F. J. Ibarrola* and R. D. Spies

5.2 Optimization and Computational Mechanics

Comprehensive comparison of performance of elite-induced evolutionary algorithm for multi-objective analysis ................................................................. 1191
Y. Shimizu* and R. Nakamura

Thermal-economic design of a micro-gas turbine CHP system using a multi-objective optimization approach ................................................................. 1203
A. C. Ferreira*, A. I. F. Vaz, S. F. Teixeira and J. C. Teixeira

5.3 Robust, Performance-based and Reliability-based Optimization under Uncertainty

Logistics of transition from working stress to limit states design in foundations engineering ................................................................. 1215
K. G. Ashkinadze*

A general approach for decoupling RBDO problems ......................................................... 1223
A. J. Torii, R. H. Lopez* and L. F. F. Miguel

Effect of expected consequences of failure on optimal structural design ......... 1235
T. G. Pedrosa and A. T. Beck*

Mechanical and reliability analyses of reinforced concrete structures subjected to corrosive effects caused by chloride ions penetration ........................................ 1245
G. P. Pellizzer*, E. D. Leonel and C. G. Nogueira

Application of intervening variables in a bilinear approach for fuzzy structural analysis ...... 1257
C. A. Pérez* and M. A. Valdebenito

Surrogate modelling techniques in solution of structural reliability problems .......... 1266
H. M. Kroetz* and A. T. Beck

SFEM using a volumetric-deviatoric split of the elasticity tensor ................................ 1277
I. Caylak*, R. Mahnken and A. Dridger

5.4 Uncertainty Modeling and Quantification

Keynote Lecturer: Marta B. Rosales

Dynamic behavior of timber beams with uncertain properties ................................. 1286
D. A. García, M. B. Rosales* and R. Sampaio

Parametric probabilistic analysis in the uncertain dynamics of magneto-electro-elastic beams resting on elastic foundations ................................. 1298
M. T. Piovan*, J. F. Olmedo and R. Sampaio

Dynamic analysis of a guyed mast with uncertainties on the stiffness and the nominal wind velocity ................................................................. 1310
J. S. Ballaben*, M. B. Rosales and R. Sampaio
5.5 Other Optimization Problems
Thermomechanical Multiobjective Bidirectional Evolutionary Structural Optimization using weighted sum method for mean compliance and heat conduction problem .......... 1331
S. Q. Rodríguez* and R. Pavanello

Integrating evolutionary optimization, statistical classification and analysis of reliability, costs and technical losses into a computational system for optimal design and diagnosis of electric power distribution grids .......................................................... 1344

Differential evolution applied to mooring optimization of offshore floating systems based on an integrated design methodology .......................................................... 1356

On the performance of the Harmony Search algorithm in the optimization of laminated composite plates .......................................................... 1368
F. S. Almeida*

Parallel genetic algorithms for the design of MIMO control systems ...................... 1380
P. P. Oteiza, D. A. Rodríguez and N. B. Brignole*

6 - Computational Acoustics

6.1 High Order Methods with Applications to Aerodynamics and Aeroacoustics
Preconditioned implicit Runge-Kutta schemes for unsteady simulations of low Mach number compressible flows .......................................................... 1387
L. S. D. B. Alves*

Numerical investigation of turbulent channel flows using a semi-spectral code ........ 1399
L. A. C. A. Schiavo, A. B. Jesus, J. L. F. Azevedo* and W. R. Wolf

A numerical study of the effects of trailing edge bluntness on airfoil tonal noise .......... 1411
W. Arias-Ramírez and W. R. Wolf*

On the development of large eddy simulation tools for compressible jet flow configurations........ 1423
C. Junqueira-Junior, S. Yamouni* and J. L. F. Azevedo

7 - Industrial Applications

7.1 Aerospace Technology
Flight dynamics parameter estimation of a rotary wing aircraft using the output error minimization with natural and meta heuristic methods .......................................... 1435
I. Y. Sumida, H. F. de Campos Velho, E. F. P. Luz, R. V. Cruz and L. C. S. Goes*
Steady and unsteady analysis of aerodynamics wing sections at ultra-low Reynolds numbers (Re < 10000) .................................................. 1446
D. P. Antonelli*, C. G. Sacco and J. P. Tamagno

Controlable planar nozzles for micro thrusters .................................................. 1459
P. H. Pedreira*, J. R. Lauretta and S. D'hers

Application of Partially-Averaged Navier Stokes simulations for bluff body aerodynamics .... 1468
S. Krajnovic* and M. Mirzaei

7.2 Industrial Applications of Computational Mechanics

**Semi-Plenary Lecturer: Eduardo N. Dvorkin**
Structural models and mechanical tests in the development of a communications spacecraft .......................................................... 1480
S. Tempone, M. Franceschini and E. N. Dvorkin*

**Keynote Lecturer: Rita Toscano**
Computational Modelling: a tool to add economic value to the industrial production ..... 1497
R. G. Toscano*, M. Goldschmit, S. Tempone and E. N. Dvorkin

Numerical simulation applied to the tire industry ....................................................... 1504
J. Kuster*, B. A. Toscano, G. N. Curtosi and T. Arechaga

A review on the nonlinear dynamics of pendulum systems for energy harvesting from ocean waves ........................................ 1516
F. E. Dotti*, F. Reguera and S. P. Machado

Analysis of self-balanced thermal stresses in long flat rectangular isotropic plates ...... 1530
T. S. Mendonça*, G. O. Ribeiro and M. A. Vecchi

Numerical simulation for identification of free spans in submarine pipelines .......... 1542
B. M. Jacovazzo, B. F. Monteiro, M. H. A. Lima Jr., J. S. Balico, F. N. Corrêa* and B. P. Jacob

On the seismic response of fractures and induced anisotropy in fluid-saturated porous media ................................................................. 1554
R. Martínez Corredor, J. E. Santos, P. M. Gauzellino* and J. M. Carcione

An harmonic 1D-element for nonlinear analysis of axisymmetric structures: the case of hot rolling ......................................................... 1566
D. Benasciutti, F. De Bona* and M. Gh. Munteanu

7.3 Modeling and Simulation of Process and Manufacturing

On using Petri nets to model and simulate processes with uncertainty .................. 1578
C. M. Ribeiro*, A. C. Lisboa, C. A. Maia and R. R. Saldanha

Reverse engineering and forming simulation of indoor of an austenitic dishwasher ...... 1587
J. C. Cisneros*, J. L. Macías and E. Royo

Non-ionizing radiation mapping based on Geographical Information Systems .......... 1598
O. Valencia, J. Huiza, M. Pohl and E. Amaya*

Application of numerical simulation to reduce the flash of a hot forged automotive piece .......................................................... 1610
G. Abate*, D. Perez, F. Riu and D. Martinez Krahmer
CONTRIBUTIONS
A PENALTY APPROACH TO OBTAIN LOWER BOUND BUCKLING LOADS OF IMPERFECTION-SENSITIVE SHELLS

LUIS A. GODOY¹, ROSSANA C. JACA², EDUARDO M. SOSA³ AND FERNANDO G. FLORES⁴

¹ Institute for Advanced Studies in Engineering and Technology, CONICET-UNC, FCEFYN, Universidad Nacional de Córdoba, P. O. Box 916, Córdoba, Argentina
lgodoy@efn.uncor.edu

² Civil Engineering Department, Facultad de Ingeniería, Universidad Nacional del Comahue Buenos Aires 1400, Neuquén, Argentina
rossana.jaca@fain.uncoma.edu.ar

³ Mechanical and Aerospace Engineering Department, West Virginia University Morgantown, WV 26506-6103, USA
eduardo.sosa@mail.wvu.edu

⁴ Institute for Advanced Studies in Engineering and Technology, CONICET-UNC, and Departamento de Estructuras, FCEFYN, Universidad Nacional de Córdoba, P. O. Box 916, Cordoba, Argentina
fflores@efn.uncor.edu

Key words: Buckling, Finite elements, Reduced Energy Method, Stability, Shells.

Abstract. A strategy, known as Reduced Stiffness (or Reduced Energy) Method, in which selected energy components are eliminated from the analysis to account for mode interaction and imperfection-sensitivity in a simplified way, has been developed by Croll and co-workers since the 1980s. This physical interpretation allows the formulation as an eigenvalue problem, in which the critical loads are lower bounds to experiments and to nonlinear incremental analysis. This paper considers the computational implementation of both reduced stiffness and reduced energy approaches to the buckling of shell structures. The structural configurations of interest in this work are cylindrical storage tanks with or without a roof. The reduced stiffness approach has been implemented in a special purpose finite element code for shells of revolution, while the reduced energy methodology was implemented in a general purpose code. The present results are compared with geometrically nonlinear analysis including shape imperfections. Links between this methodology and the Reduced Integration Technique employed in the analysis of plates/shells are discussed. Difficulties in extending the methodologies to complex problems in engineering practice are highlighted.

1 INTRODUCTION

Attempts have been made to develop simple computational tools to provide acceptable solutions to shell buckling problems and avoid gross errors in solutions without the need of having knowledge of the complete arsenal of shell buckling theory. James G. A. Croll promoted the use of a simple technique based on a reduced version of the energy (or the
stiffness) of the shell, in which only an eigenvalue problem needs to be solved. This paper discusses ways to implement such methodology in complex engineering problems using finite element codes.

The European approach to the analysis of shell buckling problems [1] using finite element tools includes GMNIA (Geometrically and Material Nonlinear Analysis with Imperfections) as the “best” estimate of buckling capacity; GNIA, Geometric Nonlinear Analysis with Imperfections; MNA, Material Nonlinear Analysis; and LBA, Linear Bifurcation Analysis. An intermediate method is recommended as a “less onerous” approach, which is based on a combination of Linear Bifurcation Analysis (LBA) and Material Nonlinear Analysis (MNA). The recommended approach requires design curves that need to be established for each geometric and load configuration, and it takes the form of elasto-plastic interaction curves. The parameters of such curves should be obtained from a number of full GMNIA, and once the curves are constructed for a specific class they can be used by performing LBA and MNA studies for a case. The non-specialist engineer who does not have the curves for his own problem is therefore lost since the start. A specialist engineer, on the other hand, needs to spend time and effort to develop the tools before using them.

The American approach is different: the loads are carefully specified, such as in the ASCE provisions [2], but the engineer is left to decide what type of analysis is suited for each case. Of course, this is not a job for the non-specialist engineer, and there are frequently cases in which someone mixes concepts and approaches to yield non-sense solutions. This has to do with legal responsibilities/insurance issues which in turn yield different philosophies for design and forensic analysis; however, it is not our aim here to discuss the merits of each regulatory approach.

The question of what is “onerous” in computational mechanics, as is the concern of the European Committee for shell buckling [1], has considerably changed over the last decades. In 2015 the “onerous” part of the job is understanding the physics of the problem and conceptually modelling the case in hand.

We shall not refer here to shell buckling problems in general terms, but attention will be restricted to the slender shells employed in the fabrication of tanks for the oil industry. Because those are very thin shells, with slenderness between $1500 < R/t < 2000$, they tend to buckle in the elastic range and plasticity develops only in the post-buckling states. There are also functional requirements that need to be considered, as in any case in industry: tanks usually have internal floating roofs that float on top of the oil or fuel, and large buckling deflections are enough to block the floating mechanism, with the consequence that the structure has to be out-of-service until it is repaired. Regarding the nature of imperfections, there are basically two types: global (due to errors in fabrication or damage under previous loads) and local (welding defects). Global imperfections usually follow eigenvalue-affine modes (i.e. modes associated with the lowest eigenvalues in LBA).

2 LBA AND GNIA METHODOLOGIES

2.1 Classical bifurcation analysis

In the classical form, a stability analysis is limited to the solution of an eigenvalue problem, also known as LBA, requires the computation of a linear equilibrium path as
\[ \mathbf{K} \mathbf{a}^F + \lambda \mathbf{P} = 0 \]  

(1)
in which the linear stiffness matrix of a shell has membrane and bending components,

\[ \mathbf{K} = \mathbf{K}^m + \mathbf{K}^b \]  

(2)

Solution of the linear equilibrium problem at a load level \( \lambda = 1 \) yields the displacements along the fundamental equilibrium path \( \mathbf{a}^F \), from which the stress resultants \( \mathbf{N}^F \) are obtained. This is an ingredient of the initial stress (or load-geometry) matrix \( \mathbf{K}^G \) \[3\]

\[ \mathbf{K}^G = \int \mathbf{N}^F \beta^T dA \]  

(3)
where \( \beta \) is the rotation vector. In LBA, an eigenvalue problem is solved (Equation 4) where the lowest eigenvalue \( \lambda^C \) is the classical critical load and \( \Phi^C \) is the associated eigenvector or critical mode.

\[ \left[ \left( \mathbf{K}^m + \mathbf{K}^b \right) - \lambda^C \mathbf{K}^G \right] \Phi^C = 0 \]  

(4)

### 2.2 Geometrically Nonlinear Analysis with Imperfections, GNIA

Computational evaluation of the nonlinear equilibrium path is now a simple task using commercial software like ABAQUS \[4\]. Benchmarks for vertical aboveground tanks have been developed by the authors in previous works and are employed in this one. Four tanks opened at the top were presented in Godoy and Flores \[5\], with different geometries in such way that they have significantly different imperfection sensitivity. Three tanks with conical roof were introduced by Sosa et al. \[6\] with differences in the height of the cylindrical part; they serve to discuss shells which exhibit the same imperfection-sensitivity although they have differences in their geometry.

To illustrate the nature of results, consider an opened top tank (a cantilever cylinder with step-wise variable thickness) under uniform pressure (Figure 1). In all cases reported in this paper, API 650 \[7\] regulations have been used to evaluate the thickness.

![Figure 1](image-url)  

**Figure 1**: Cantilever cylindrical tank with step-wise variable thickness. \( D = 30.48\text{m}, H = 13.10\text{m}, t = 7.9\text{mm} \) at the top three courses. (a) Non-linear paths for different level of imperfections obtained from GNIA, (b) Imperfection sensitivity curve \[6\].
GNIA has been performed in the case of Figure 1, in which the path and the imperfection-sensitivity are shown. At point C in Figure 1, the maximum in the path is $\lambda^{\text{max}} = 0.76$; this is the largest value of imperfection amplitude $\xi$ for which a maximum is recorded, and is identified as a lower bound computed with GNIA.

This case shows moderate imperfection-sensitivity, with a reduction between the classical critical load (LBA solution) and the maximum in the path for $\xi = t_{\text{min}}$ of approximately 25%. This approach is considered onerous in the ECCS [1] regulations, so that simpler methods have been envisaged to tackle this problem.

3 REDUCED STIFFNESS ANALYSIS (RSA)

The procedures for implementation of Reduced Energy Method (REM) and the Reduced Stiffness Method (RSM) have been reviewed by Croll [8] and Sosa et al. [6], just to cite a couple of references. The necessary background to the energy approach to investigate stability of elastic systems was originally developed by Koiter for continuous systems [9], followed by developments for discrete systems [10], [11].

3.1 Methodology

The RSA (also the REA presented in the next section) starts by evaluating the energy components of the shell in the classical eigenmodes, including membrane and bending energy contributions and load potential. Depending on the shell and load considered, some of these energy contributions are positive and others are negative. Positive energy components contribute to the stability of the shell, while negative components tend to de-stabilize the shell. The main assumption in the RSA is that positive stabilizing components may be lost in the buckling process due to coupling between geometric imperfections and nonlinear effects.

Under lateral pressure, it has been shown that membrane components are eroded in cylindrical shells [8]. Thus, reducing the membrane energy contribution is carried out in the RSA, and in the limit, as the membrane contribution is eliminated, a lower bound is reached. This approach has been extensively validated by comparison with experiments and with GNIA computations, at least for uniform pressure cases.

In the RSA membrane components in the linear stiffness matrix $K$ are selectively eliminated, but this may be more conveniently written in terms of a penalty approach using a penalty parameter $\alpha$, with values between zero and one. Penalty parameters have been used in finite element analysis for some time (see for example Zienkiewicz et al., [12], pp. 88). The modified LBA becomes:

$$\left[ (\alpha K^m + K^b) - \lambda' K^G \right] \Phi' = 0$$

The solution of this modified eigenvalue problem is identified as $\lambda'$, $\Phi'$, which depend on the adopted $\alpha$. As $\alpha \to 0$, then the problem tends to the classical form of the RSA. It has been shown that a lower bound is obtained for small values of $\alpha$, not necessarily zero [13,14].

Notice that the equilibrium condition yielding $\mathbf{N}^F$ is not affected by $\alpha$. A linear equilibrium path $P_1$ serves to identify $\lambda^C$ and also $\lambda'$ using RSA. But if $\alpha$ was also employed to compute an equilibrium path $P_2$, then RSA would yield a different $\lambda''$, not being a lower bound to experimental or GNIA results.
The knock-down factor $\eta$ is usually employed to normalize results in the form

$$\lambda' = \eta \lambda^C$$  \hspace{1cm} (6)

Because $\Phi'$ has not been limited to coincide with $\Phi^C$, the eigenmodes may change from the classical to the RSM, thus providing more flexibility to the analysis. The drawback is that this can only be implemented in a special purpose code by modification of the formulation [15], but it has not yet been possible to implement it in general purpose codes like ABAQUS [4].

3.2 Application of RSA in cases for which $\Phi' = \Phi^C$

Implementation of the RSA has been done in this case using a special purpose finite element code ALREF originally developed by Flores [15]. This is a semi-analytical finite element model which is capable of performing LBA as well as initial post-critical behavior [16]. Because this is an in-house code, changes in the computations can be introduced to reduce the membrane contribution to the stiffness of the shell.

To illustrate the use of this methodology, the open top tank already solved in Figure 1, is considered.

In this case the modes satisfy the condition $\Phi' = \Phi^C$, and results are presented in Curve (1) of Figure 2. The knock-down factor using RSA is $\eta = 0.77$, whereas the lower bound using GNIA is $\eta = 0.76$. The present approach is capable of reproducing what would be obtained using a far more onerous full nonlinear analysis.

Many other cases have been solved with similar agreement of results, thus supporting the accuracy of the penalty formulation.

![Figure 2: Open top tanks, results of penalty approach to RSA: (a) open top tank of Figure 1 under internal pressure; (b) open top tank with \(D = 14m, H = 3.5m, \) uniform \(t = 4mm, \) under wind. Results computed with ALREF. Penalty parameter is plotted as \((1-\alpha)\). Adapted from Jaca et al. [13,14].](image)

3.3 Application of RSA in cases for which $\Phi' \neq \Phi^C$

Shell problems under wind loading are typical cases in which the classical eigenmodes does not coincide with that obtained from the RSA. Notice that all analytical results published by Croll and co-workers are for axisymmetric loads, except for the cooling tower studies of Zintillis and Croll [17], in which wind pressures were taken into account in a simplified way.

Again, ALREF [15] has been used to perform the computations, with changes in the code to account for reductions in the membrane contribution. The specific case reported here is a shell opened at the top, with \(r/t = 1750 \) and \(H/D = 0.25 \). Wind was modeled according to
ASCE 7 [2]. To compare results, a full GNIA study was performed. Results are plotted in Curve (2), Figure 2.

In this case, there are differences in mode shape rather than in eigenvalues. Classical buckling modes present large deflections in the windward region, known as “mode attenuation” around the circumference, thus becoming a localized mode. For the leeward region of the shell there are small deflections as obtained from classical buckling studies. The mode localization affects not just the meridional but also the circumferential directions. Implementation of the RSA produces a reduction in membrane contributions of the complete shell, which may be quite different from what is seen in cases of global modes. The mode shapes obtained via RSA show mode differences, with spurious deflections at leeward. The knock-down factor $\eta = 0.55$ obtained with RSA is lower than the reduction obtained with GNIA.

An improved solution was obtained using a selective stiffness reduction, in which only the top courses (with thinner shell, where buckling occurs) are considered for the membrane reduction. This selective stiffness reduction produced a significant improvement in mode shape, becoming close to the shape identified with GNIA [18].

The cost of improving a solution in the case is high, in the sense that not only a special purpose code is required, but also the method should be applied on the part of the shell in which buckling develops.

4 REDUCED ENERGY ANALYSIS (REA)

4.1 Methodology

The Reduced Energy Analysis (REA), has been described by Croll and co-workers [8], but a slightly different formulation is given here. The quadratic form may be written in the form

$$\Phi^T \left[ (K^m + K^b) - \lambda C K^G \right] \Phi = 0$$

(7)

The penalty parameter $\alpha$ is next introduced

$$\Phi^*^T \left[ (\alpha K^m + K^b) - \lambda^* K^G \right] \Phi^* = 0$$

(8)

Both, classical and modified quadratic forms may be written as:

$$\Phi^T \left[ (K^m + K^b) \right] \Phi = \lambda C \Phi^T K^G \Phi^C \quad \text{and} \quad \Phi^*^T \left[ (\alpha K^m + K^b) \right] \Phi^* = \lambda^* \Phi^*^T K^G \Phi^*$$

(9)

Assuming that the eigenmodes are the same in both problems, $\Phi^* = \Phi^C$, one gets

$$\frac{\lambda^*}{\lambda} \Phi^C = \Phi^T \left[ (\alpha K^m + K^b) \right] \Phi^C$$

(10)

In the Theory of Elastic Stability, the following notation is usually employed

$$U^{2m} = \Phi^T K^m \Phi^C \quad \text{and} \quad U^{2b} = \Phi^T K^b \Phi^C$$

(11)
in which $U^{2m}$ and $U^{2b}$ are energy components in the critical mode. Only the eigenvalue $\lambda^*$ is left as an unknown, and it may be computed for a given value of $\alpha$.

$$\lambda^* = \eta \lambda^C$$

One of the limitations of implementing this REA in a general purpose code like ABAQUS is that it does not allow for separate computation of membrane and bending energy components, which are merged in any finite element package.

### 4.2 Metal shells modeled as a composite material

Sosa et al. [6] showed a simple way to implement the REA by modelling the isotropic metal shell as a composite material. Thus, by assuming a more complex material model, one may employ a constitutive matrix of a composite as in Classical Lamination Theory [19].

$$\begin{bmatrix} N \\ M \end{bmatrix} = \begin{bmatrix} A & B \\ B^T & D \end{bmatrix} \begin{bmatrix} \varepsilon \\ \kappa \end{bmatrix}$$

(A is the membrane sub-matrix, whereas the bending sub-matrix is $D$, given as)

$$A = \begin{bmatrix} A_{11} & A_{12} & 0 \\ A_{21} & A_{22} & 0 \\ 0 & 0 & A_{33} \end{bmatrix} \quad D = \begin{bmatrix} D_{11} & D_{12} & 0 \\ D_{21} & D_{22} & 0 \\ 0 & 0 & D_{33} \end{bmatrix}$$

(B is the bending-extension coupling matrix, which is a null matrix for symmetric laminates. The REA is implemented by adopting a reduction in all or selected membrane components of $A$ with a penalty parameter, $\alpha^A$. Notice that by affecting the membrane energy of the shell, a different shell theory is being assumed.

The computations follow this sequence:

- The classical eigenvalue problem is solved, leading to $\lambda^C$, $\Phi^C$.
- The mode shape $\Phi^C$ is imposed as a displacement field in a linear static analysis of the shell as a composite with $\alpha = 1$. The strain energy is computed. This is $(U^{2m} + U^{2b})$.
- The mode shape $\Phi^C$ is imposed as a displacement field in a linear static analysis of the shell as a composite with $\alpha \neq 1$. The strain energy is computed. This is $(\alpha U^{2m} + U^{2b})$.
- The reduced load is calculated as the ratio between reduced and full energy values at the critical state.

Eight elements in ABAQUS have been employed to perform computations, and it was found that element STRI3, a fully integrated triangular element originally developed by Batoz et al. [20], showed the best performance by converging to the analytical solution of REM in a benchmark case [6]. Similar performances were found with other commercial software.

### 4.3 Application of REA in cases for which $\Phi^* = \Phi^C$

The equivalent composite material has been employed to solve tanks with variable thickness which are opened at the top, and tanks with a conical roof on top, both under uniform pressure. Results were compared with GNIA obtained lower bounds and showed excellent results.
Figure 3 shows the penalty approach for tanks with conical roof, with identical geometries but different heights $H$. Imperfection sensitivity in all three cases is the same because the tanks buckle at the top in a local mode. Results with GNIA lower bounds provided $\eta = 0.73$ (case $H/D = 0.4$). It was shown that the assumption $\Phi^* = \Phi$ was adequate in this class of problems, i.e. there was not a change in mode shape from the classical to the REA computations. The REA (case $H/D = 0.4$) provided a value slightly higher than GNIA.

Values of $\eta$ are plotted in Figure 3(a) for values of $1/\alpha$ (rather than $\alpha$, as in the previous section on RSA) in order to distinguish results in the zone close to full reduction of membrane energy. One of the differences shown in Figure 3(a) is that not all three cases considered tend to the same value as the penalty parameter $\alpha$ tends to zero: the shorter tank tends to a higher value than the other two cases.

Figure 3: Tanks with conical roof. $D = 30.48\, \text{m}$, $t = 7.9\, \text{mm}$ at the three top courses. REA computed with ABAQUS: Knock-down factor calculated with first eigenmode. (a) Uniform pressure [6]; (b) wind pressure [21].

Results under uniform pressure were encouraging and showed a possible way to implement the REA in finite element software like ABAQUS, without introducing any modification in the code, simply by using its capabilities.

4.4 Application of REA in cases for which $\Phi^* \neq \Phi^C$

Wind loaded shells display such behavior that one cannot assume that the classical eigenmode is also the same deflection pattern associated with the penalty formulation. Sosa and Godoy [21] showed that there are significant differences in the number of waves and wave pattern around the circumference.

The only previous works in which some form of wind pressures were reported by Zintillis and Croll [17] for hyperboloids of revolution (cooling tower shells). However, buckling using REA was estimated for axisymmetric pressure and the results were extended to wind by adopting a “Worst Stress Meridian” (WSM) and a “Worst Pressure Meridian” (WPM) methods. Comparisons with wind tunnel tests showed differences in the order of 30%, and a number of reasons may serve to explain such differences.

The case reported in Figure 3(b) are tanks with conical roof, with the same dimensions as those reported in Figure 3(a) but here solved under wind [21]. Using GNIA, the lower bound is $\eta = 0.6$. The REA, on the other hand, provided much higher values in the range $\eta = 0.76$ to 0.79, and with variations among them, as seen in Figure 3(b).

The second case reported here are open tanks with differences in geometric characteristics, as adopted from [5], under wind. The cases have different imperfection-sensitivity. GNIA results yield $\eta = 0.60$ for the tank with $H/D = 0.17$ (tank M1), with $\eta = 0.95$ for the tank with
H/D = 1 (tank M4), as shown in Figure 4. The significant differences in $\eta$ that should be expected from these quite different geometries (as reflected by GNIA) are not seen in the REA, in which all cases seem to be very close to each other. Because the eigenmode is not computed in REA, the classical mode is assumed, and this is not adequate in this case.

Figure 4: Open-top tanks. REA results: Imperfection-sensitivity should depend on shell geometry

5 SELECTIVE INTEGRATION AS AN ALTERNATIVE TO PENALTY FORMULATION

For a number of years, finite element formulations for plates and shells have employed some form of reduced integration of the element stiffness matrix in order to avoid different types of locking (shear, membrane, volumetric) (see, for example, Zienkiewicz et al. [12]). Locking appears due to the impossibility of the interpolated displacements to simultaneously satisfy certain kinematic constraints. In selective integration, a low order numerical integration rule is applied to some parts of a matrix or sub-matrix in a problem. It has been shown that this improves results with respect to “exact” integration rules because it avoids locking of an element.

In fact, reduced integration for solving locking problems has been replaced in most cases by other type of sound basis formulations, like assumed natural strains or enhanced assumed strain methods. Nowadays reduced integration is mainly used in explicit codes to reduce the number of computations and simulation time, i.e. improve code performance. Results in this case are expected to agree (to a certain degree) with full integration at a quite lower computational cost, so the target is not a stiffness reduction. Besides under-integration requires some sort of stabilization to avoid the appearance of spurious zero-energy modes (hourglass or checkerboard modes).

The question arises as to what is the relation between reduced/selective integration methods and RSA. Can RSA be implemented in practice via selective integration? In our shell buckling problem, only the membrane part in the LBA would need to be integrated with a lower order.

To illustrate the possibilities of selective integration as a tool in search for lower bound buckling loads, consider a simple one degree of freedom problem given by

$$\left( k^m + k^b \right) - \lambda C^G k^G = 0$$

(15)

For simplicity, assume that $k^b$ is a fraction of $k^m$ in the form $k^b = \beta k^m$, with $\beta < 1$; $k^m$ is assumed in the simple form
where \(-1 < x < 1\) is a normalized domain.

In the RSA, the problem is posed as

\[
\left(\frac{1}{\alpha}k^m + k^b\right) - \lambda^* k^G = 0
\]  

(17)

For this simple system, \(\lambda^*\) is obtained by substitution of \(k^m\) and performing analytical integration

\[
\lambda^* = \left(\frac{1}{\alpha} + \beta\right) k^m = 5.4857 \left(\frac{1}{\alpha} + \beta\right) \frac{\gamma}{k^G}
\]  

(18)

Selective integration is next carried out

\[
\left(\bar{k}^m + \beta k^m\right) - \lambda^* k^G = 0 \quad \text{leading to} \quad \bar{\lambda} = \left(2 + 5.4857 \beta\right) \frac{\gamma}{k^G}
\]  

(19)

For \(\beta = 0.3\), the condition \(\bar{\lambda} = \lambda^*\) is satisfied only for \(\alpha = 2.74\). The equivalence in terms of this value of \(\alpha\) also holds for other values of \(\beta\). Results are illustrated in Figure 5.

![Figure 5: RSA and selective integration for a simple 1 DOF system.](image)

Although some reduction is achieved by using low-order integration, this equivalence occurs for a value which is of 65% of the expected reduction as expected from the lower bound RSA. The conclusion from this simplified example is that reduced integration is expected to produce a significant reduction in the buckling load, but not as much as is predicted by a more specialized RSA.

6 CONCLUSIONS

New computational methodologies to implement lower bound buckling estimates for imperfection-sensitive shells have been discussed in this paper. The methodologies reduce the membrane contribution in the classical eigenvalue problem by means of a penalty parameter and have been implemented in a special purpose finite element code (ALREF) via Reduced Stiffness, and in a general purpose finite element code (ABAQUS) via Reduced Energy
Analysis. In all cases, the aim is to estimate lower bounds while at the same time adequately representing the buckling mode. Attention to this last point has not been given in previous research, and it emerges here because of the need to compute new mode shapes under wind pressures.

Some conclusions may be drawn as follows:

- REA and RSA are both based on the same philosophy, but with different assumptions. RSA does not assume the mode shape at lower bound, and calculates it by solving a modified eigenvalue problem. REA assumes that the lower bound mode is the same as the classical mode.
- Both REA and RSA yield good lower bound estimates for buckling load under uniform pressure, for which the mode in the classical LBA is the same as in the reduced methodology.
- The use of RSA requires modification of a finite element code, and this can be achieved in special purpose codes, not in general purpose ones. This becomes an obstacle to the adoption of this methodology in engineering practice.
- Modeling the isotropic metal shell as an orthotropic composite is a simple way to implement REA using a general purpose code. This proved to be a simple procedure.
- It has been shown that extension of RSA and REA to cases in which the mode changes between classical and lower bound solutions, is not a simple task. Our results indicate that a REA implementation was not able to identify significant differences in imperfection-sensitivity benchmarks for opened top tanks under wind. The problems are less severe in RSA, with good estimates in lower bound buckling loads, but difficulties arise in the identification of the correct eigenvectors.
- Use of a selective integration technique instead of a penalty methodology does not seem to be a promising way: less flexibility is introduced because it is equivalent to using a fixed penalty parameter.

The current status of REA and RSA are not sufficient to become a viable general purpose tool to shell buckling problems in practical engineering. They are, however, powerful tools to enhance understanding of the physics of shell buckling and will surely enlighten further research.

**Acknowledgements:** The authors are indebted to Prof. James G. A. Croll, who developed the lower bound methodology discussed in this work. LAG and FGF thank the support received through grants from the National University of Cordoba (SECYT-UNC) and CONICET (PIP112-201201-00126-CO). RCJ thanks the support during this research received through grants from the National University of Comahue (SECYT-UNCo). LAG and FGF are members of the Research Staff of CONICET.

**REFERENCES**


BUCKLING OF COMPOSITE CROSS-PLY SANDWICH PANELS USING DIFFERENT FINITE ELEMENT APPROXIMATIONS

FERNANDO G. FLORES¹

¹Departamento de Estructuras, Universidad Nacional de Córdoba And CONICET
Casilla de Correo 916, 5000 Córdoba, Argentina
fflores@efn.uncor.edu, http://www.efn.uncor.edu

Key words: Sandwich Composites, Buckling Loads, Refined Zigzag

Abstract. The use of composite materials in different areas of engineering has grown markedly in recent years. Parallelly various techniques for modeling macroscopic laminate have been developed. In this paper the buckling loads of sandwich panels with cross-ply sandwich sections have been obtained numerically using finite elements based on different shell theories, i.e. classical lamination theory, first order transverse shear theory, refined zigzag theory. The elements considered are shell-type that have between 3 and 7 degrees of freedom per node while this numer does not depend on the number of layers or the characteristics of the section. In comparisons numerical models with solid elements, involving a quite larger computational cost, are used as reference results. Examples discussed include flat panels subjected to in-plane loads and cylindrical panels subjected to in-plane loads or normal pressure. The study shows that buckling loads and critical modes may vary significantly depending on the approximation used.

1 INTRODUCTION

The use of laminated composite materials has grown dramatically in the last 30 years. Due to its high stiffness and excellent strength/weight ratio, their use is becoming more common in both aeronautical and land vehicles, as well as parts of civil structures. A particular case are sandwich sections in which the material comprising the outer layers is much stiffer and stronger than the central portion or core of the section. Such sections are best suited to resist bending due to its high flexural stiffness. In turn these sections may be subjected to compressive stresses which can lead to instability and buckling. The correct computation of buckling loads depends on a proper evaluation of the section properties including the transverse shear stiffness as it can be remarkably low in sandwich sections. The analysis of a composite laminate requires to include various simplifying assumptions to establish a manageable numerical model. Micro-mechanical models, i.e. where the
different material components are distinguished, may be used for studying the behavior of a single lamina that allows to evaluate the macroscopic properties or to establish the equivalent properties as if it were a single orthotropic material. In turn these equivalent properties are restricted to the range of elastic strains. A laminate composed of multiple laminae oriented in different directions in turn requires hypothesis to establish macroscopic properties based on the components, orientation and position of each lamina. This results in different analysis strategies with different accuracy and computational cost.

The simplest approach that maintain the problem as bidimensional is the classical laminated theory (CLT), where the Kirchhoff assumptions for isotropic thin shell are adopted so that the influence of the transverse shear is neglected in the deformation. A major improvement is to use the Mindlin-Reissner approach or first order transverse shear deformation theory (FSDT) in which the influence of transverse shear flexibility of the laminate is included. The evaluation of the section flexibility is generally based on energy criteria similar to the classical form used in homogeneous sections that leads to factor \( k = \frac{5}{6} \) that affects the cross-sectional area in the shear stiffness. These factors can be very low in sandwich sections due to the high stiffness ratio between the outer layers and the core which is usually made of a material with elastic modulus several orders of magnitude smaller than the corresponding to the outer layers. These factors are not easily computed and may be appropriate in some cases and not in others. Both approaches mentioned above have significant limitations, the first because it excludes the shear deformations and the second because the shear strain is assumed constant and therefore violates the equilibrium conditions on the boundary and continuity of shear stresses.

There is also a number of higher order theories (HOT) that seeks to properly approximate the transverse behavior and improve the quality of the predictions. A detailed discussion can be seen in [1] and references cited therein. Among the innovative approaches is the refined theory of Zigzag (RZT) [2] that, although does not satisfy the continuity of interlaminar shear stresses, has shown a very good predictive ability. The through-the-thickness integration of local equilibrium equations can be used to recover the shear stresses with very good approximation.

The most reliable approximations are obtained using discretizations that depends on the number of laminate layers, either using 3D elements or layers-wise approximation. These have the major disadvantage that the number of degrees of freedom may become unmanageable depending on the structural shape and the number of layers.

This paper attempts to assess the accuracy of different approaches in determining buckling loads of flat and cylindrical panels under compression. Only sandwich sections have been considered where the transverse shear behavior is most relevant. The approaches considered are CLT, FSDT and RZT. Comparisons are made against three-dimensional (3D) solid models. In all cases clamped or symmetry boundary conditions have been considered that are the simplest to simulate in the case of 3D models and allow a clear comparison with shell approximations.
2 NUMERICAL MODELS CONSIDERED

Numeric evaluations below has been performed using an in-house implicit finite element program for the analysis of elastic structures with large displacements and rotations. The program computes buckling loads and modes solving an eigenvalue problem of the form

\[ \mathbf{K}_0 + \lambda_c (\mathbf{K}_P - \mathbf{K}_0) \phi = 0 \]  

where:
- \( \mathbf{K}_0 \): is the initial stiffness matrix
- \( \mathbf{K}_P \): is the tangent stiffness matrix perturbing the configuration \( \hat{x} = x_0 + h\mathbf{u}_T \)
- \( \mathbf{u}_T \): is the tangent direction to the primary path \( \mathbf{u}_T = \mathbf{K}_0^{-1}f \)
- \( f \): is the external load vector
- \( \lambda_c \): is the buckling or critical load factor
- \( \phi \): is the buckling or critical mode

The finite elements used for the different approximations are:

- Classical laminated theory (CLT), a rotation-free linear triangle is used [3]
- First order shear deformation theory (FSDT), a bi-linear quadrilateral [4] with standard shear factor of isotropic materials 5/6 is used
- First order shear deformation theory plus shear correction factor (FSDT+SF), uses the same quadrilateral above and the shear factors computed as suggested in [5]
- Refined zigzag theory (RZT), a bi-linear quadrilateral as described in [6] is used.
- Three dimensional (3D), a tri-linear 8-node brick element with an improvement in transverse shear for shell models [7] is used.

3 UNIFORMLY COMPRESSED SQUARE CCCC PLATE

A square plate (side \( a = 1 \)) clamped along the four sides under uniform pressure in one direction (\( x \)) and free to uniformly displace in the orthogonal direction (\( y \)) is firstly considered. The numerical model of the plate includes one quarter only of the total geometry, and the symmetry lines coincide with the Cartesian axes. The plate is allowed to uniformly expand in \( y \) direction, this implies for the primary path a constant strain for both Cartesian directions at every point of the plate, with a) a constant compression force \( N_x \) and b) a null force \( N_y = 0 \). For the 3D model, due to the symmetry of the cross sections considered, symmetric transverse displacements \( u_z \) are imposed, i.e. symmetric points have the same displacement but with sign reversed. This allows the free contraction/expansion of the laminate without transverse normal stresses as in shell theories in order to obtain a better correlation. The discretization of one quarter of the plates includes 32 uniformly spaced elements per side.
3.1 Section 1

An analytical solution is provided for rectangular plates in [8]. Three sandwich sections are considered there with the data shown in Table 1 where the external layers have a thickness of 0.5mm while the core thickness \( h \) may have 3 different values. The component materials of each part have the mechanical properties in Table 2 where the bold-faced values are provided in the Ref. [8] and the other values have been estimated to be used in this paper. The shear correction factors results \((k_x, k_y) = (0.591, 0.405)\) for \( h = 10 \) and \((k_x, k_y) = (0.873, 0.762)\) for \( h = 50 \).

Table 1: Stacking sequence of the laminates for section 1 (plate).

<table>
<thead>
<tr>
<th>Mat</th>
<th>( t ) [mm]</th>
<th>Orient.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>10.0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2: Mechanical properties for section 1 \((E_i \text{ and } G_{ij} \text{ in GPa})\)

<table>
<thead>
<tr>
<th>Mat</th>
<th>( E_1 )</th>
<th>( E_2 )</th>
<th>( E_3 )</th>
<th>( \nu_{12} )</th>
<th>( \nu_{13} )</th>
<th>( \nu_{23} )</th>
<th>( G_{12} )</th>
<th>( G_{13} )</th>
<th>( G_{23} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>54.55</td>
<td>54.55</td>
<td>15</td>
<td>0.32</td>
<td>0.32</td>
<td>0.32</td>
<td>20.67</td>
<td>3.78</td>
<td>3.78</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
<td>0.32</td>
<td>0.32</td>
<td>0.32</td>
<td>0.44</td>
<td>0.44</td>
<td>0.22</td>
</tr>
</tbody>
</table>

For the 3D model the discretization includes 2 elements at each external layer and 3 elements for the core. For the first two core thickness global buckling modes are obtained where both faces and the core displace together transversely. For \( h = 100 \) the transverse displacements of the faces in the buckling mode are opposite, i.e. the core is compressed or stretched alternatively in the transverse direction. This of course can not occur in shell models.

The Table 3 shows the buckling load for the three core thicknesses using the different approaches. The analytical values of Ref [8] are also included.

For the thinnest \( h \) the largest difference in buckling load is 3.3%. Naturally this maximum load is obtained with CLT while the FSDT gives a buckling load of 1.46% in excess of the 3D model. The RZT gives a buckling just below the 3D as also does the FSDT+SF.

Table 3: Buckling loads for section 1 [kN]

<table>
<thead>
<tr>
<th>( h ) [mm]</th>
<th>Ref.</th>
<th>3D</th>
<th>FSDT</th>
<th>FSDT+SF</th>
<th>RZT</th>
<th>CLT</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>161.6</td>
<td>164.3</td>
<td>166.7</td>
<td>164.0</td>
<td>164.0</td>
<td>169.7</td>
</tr>
<tr>
<td>50</td>
<td>3275</td>
<td>3518</td>
<td>3556</td>
<td>3543</td>
<td>3526</td>
<td>4117</td>
</tr>
<tr>
<td>100</td>
<td>11292</td>
<td>12357</td>
<td>12483</td>
<td>12598</td>
<td>16289</td>
<td></td>
</tr>
</tbody>
</table>
For $h = 50$ differences notably increases for the CLT (17%) but keep quite good for the FSDT (1.08%) compared with the 3D model. Best results are obtained with the RZT (just 0.22% above) while the FSDT+SF provides also a very good approximation (0.71%).

The analytical values are slightly below those obtained in this work. This may be due to the simplifying hypothesis used or the mechanical properties adopted in this work that were missing in the Ref.[8]. Besides that the buckling modes are similar in all the simulations with a wave in both in-plane direction (see Figure 1 for $h = 10$ and 50) as assumed in the reference.

It may be concluded that for this set of materials where the mechanical properties have a ratio of 10, the results obtained with the FSDT are in close agreement with the 3D model. On the other hand the CLT give increasing differences as the thickness increase.

### 3.2 Section 2

A second section where the orthotropic component materials have a ratio of 1000 between the mechanical properties has been considered. This is two orders of magnitude larger than in section 1. The Table 4 shows the mechanical properties of the component material while the Table 5 indicates the thickness, orientation and stacking sequences for the two laminae orientations where two total thicknesses will be used (10mm or 50mm). The core material has only transverse anisotropy while the stiff skins are made of two cross-ply layers with two possible orientations as indicated in Table 5.

<table>
<thead>
<tr>
<th>Mat</th>
<th>$E_1$</th>
<th>$E_2$</th>
<th>$E_3$</th>
<th>$\nu_{12}$</th>
<th>$\nu_{13}$</th>
<th>$\nu_{23}$</th>
<th>$G_{12}$</th>
<th>$G_{13}$</th>
<th>$G_{23}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>10</td>
<td>50</td>
<td>10</td>
<td>0.05</td>
<td>0.05</td>
<td>0.25</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>0.01</td>
<td>0.01</td>
<td>0.07585</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.0225</td>
<td>0.0225</td>
<td>0.0225</td>
</tr>
</tbody>
</table>

The shear correction factors for the first orientation were computed as $(k_x, k_y) = (0.02334, 0.02167)$; for the second orientation the factors are obtained just permuting the indexes.
Table 5: Stacking sequence of the laminates for section 2

<table>
<thead>
<tr>
<th>Mat</th>
<th>thickness [%]</th>
<th>Orient.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>90</td>
</tr>
<tr>
<td>4</td>
<td>80</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>90</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>

(1)

<table>
<thead>
<tr>
<th>Mat</th>
<th>thickness [%]</th>
<th>Orient.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>5</td>
<td>90</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>80</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>90</td>
</tr>
</tbody>
</table>

(2)

Table 6: Buckling loads for section 2 [kN]

<table>
<thead>
<tr>
<th>t[mm]</th>
<th>Orient.</th>
<th>3D</th>
<th>FSDT</th>
<th>FSDT+SF</th>
<th>RZT</th>
<th>CLT</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1</td>
<td>80.95</td>
<td>105.6</td>
<td>80.52</td>
<td>80.92</td>
<td>106.1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>77.76</td>
<td>104.0</td>
<td>77.79</td>
<td>77.76</td>
<td>104.6</td>
</tr>
<tr>
<td>50</td>
<td>1</td>
<td>1306</td>
<td>10885</td>
<td>1189</td>
<td>1313</td>
<td>12681</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1164</td>
<td>10613</td>
<td>1104</td>
<td>1170</td>
<td>12548</td>
</tr>
</tbody>
</table>

In Table 6 buckling loads for the two thicknesses considered, with the two possible stacking sequences and the five numerical models considered in this work are indicated. For the thinnest case, where transverse shear strains are expected to have less influence, it can be seen that effectively the CLT and the FSDT give very similar results. However these values are more than 30% higher than the reference 3D model which is almost coincident with the RZT. Using shear correction factors (FSDT + SF) results are also very similar to the 3D model. Bifurcation modes have in all cases a full wave in each direction similarly to the section 1 as shown in the left side of Fig. 2.

For thickness $t = 50$ mm instead the buckling mode contains two complete waves in the loading direction (right side of Fig. 2). The RZT model is in very good agreement with the 3D model in both buckling load and mode. Bifurcation loads computed with CLT and FSDT models are in this case one order of magnitude higher than those obtained with the 3D model. Furthermore as shown in the left part of Fig. 3 the bifurcation mode

![Figure 2: Buckling modes of the CCCC plate, section 2](image)
maintains a single longitudinal wave. Finally if correction factors (FSDT +SF) are used the buckling loads obtained are below the 3D model (between 91 and 95% of the 3D model) and the bifurcation modes are erroneous (see right side of Fig. 3).

4 CYLINDRICAL PANEL

In this example the performance of a fully clamped cylinder subjected alternately to external pressure and axial loads is studied. The diameter of the cylinder is 1 and the total length considered is 2. The same sections defined in Table 5 with thickness $t = 10\text{mm}$ and $t = 50\text{mm}$ are considered. The principal directions of the laminate are: 1 coincident with the cylinder parallel and 2 in the direction of the cylinder meridian. The numerical model includes only one eighth of the cylinder (1 quadrant and half the length) imposing symmetry conditions in three of its edges and the remaining is clamped. This arbitrarily restricts bifurcation modes to such symmetries but substantially alleviates numerically the 3D solid model used as a reference. Discretization with quadrilateral shell elements involves 2623 nodes and 2520 elements resulting from a structured mesh with 60 divisions in one quadrant and 42 divisions along the cylinder half. The triangular shell elements discretization has the same nodes and twice the elements. The 3D model with hexahedral solid elements has the same discretization used for mid-side surface and 7 elements across the thickness.

4.1 Results for external pressure

The critical loads obtained for the case of external pressure are indicated in Table 7. A very good correlation between the 3D model and the shell model when the RZT is used can be seen, with minor differences of 0.5% for $t = 10\text{mm}$ and a maximum of 2.2% for $t = 50\text{mm}$. In Fig 4 buckling modes as seen from the cylinder axis are shown, the first row is associated with orientation 1 and the second row with orientation 2. The number of waves is the same except for $t = 50$, orientation 2 where RZT model indicates a greater number of waves.
Table 7: Critical loads under external pressure [MPa]

<table>
<thead>
<tr>
<th>( t ) [mm]</th>
<th>Orient.</th>
<th>3D</th>
<th>FSDT</th>
<th>FSDT+SF</th>
<th>RZT</th>
<th>CLT</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1</td>
<td>0.05746</td>
<td>0.06644</td>
<td>0.05750</td>
<td>0.05774</td>
<td>0.06640</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.05217</td>
<td>0.06011</td>
<td>0.05261</td>
<td>0.05242</td>
<td>0.08031</td>
</tr>
<tr>
<td>50</td>
<td>1</td>
<td>1.156</td>
<td>3.604</td>
<td>1.128</td>
<td>1.175</td>
<td>3.730</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.016</td>
<td>3.327</td>
<td>1.047</td>
<td>1.039</td>
<td>3.433</td>
</tr>
</tbody>
</table>

Figure 4: Buckling mode under external pressure (a-b) 3D; (c-d) RZT;

The largest differences occur naturally for CLT ranging from 15% (same critical mode) and 54% (shorter wavelength) for the thin section to more than 200% for the thick section with a larger wavelength buckling mode. For the FSDT model the buckling modes matches the 3D model for the thin case but gives a smaller number of waves for the thick case, while the buckling loads exceed 3D model by 15% and more than 200% respectively. Finally if correction shear factors are used, bifurcation loads fit reasonably well to those obtained in the 3D model with differences smaller than 1% for \( t = 10 \), with a similarly mode, and about 3% for \( t = 50 \) but with a shorter wavelength.

4.2 Results for axial load

Critical loads obtained for axial load are indicated in the Table 8 where again a be a very good correlation between the solid and the RZT models can be seen, with minor differences of 0.25% for \( t = 10 \) mm and a maximum of 2.0% for \( t = 50 \) mm. The associated buckling modes are shown in Fig. 7, the first row is associated with orientation 1 and
the second with orientation 2. It can be seen that the buckling pattern is the same for \( t = 50\text{mm} \) while for thin case the buckling modes does not spread over the entire cylinder for the RZT model (may be due to some numerical convergence issues) but the pattern has the same wavelength in both cases.

Table 8: Critical loads under axial load [MN]

<table>
<thead>
<tr>
<th>( t[\text{mm}] )</th>
<th>Orient.</th>
<th>3D</th>
<th>FSDT</th>
<th>FSDT+SF</th>
<th>RZT</th>
<th>CLT</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1</td>
<td>0.2316</td>
<td>0.3372</td>
<td>0.2223</td>
<td>0.2322</td>
<td>0.3448</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.2630</td>
<td>0.3378</td>
<td>0.2392</td>
<td>0.2636</td>
<td>0.3473</td>
</tr>
<tr>
<td>50</td>
<td>1</td>
<td>1.740</td>
<td>9.134</td>
<td>1.113</td>
<td>1.773</td>
<td>9.708</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.887</td>
<td>9.445</td>
<td>1.197</td>
<td>1.924</td>
<td>10.082</td>
</tr>
</tbody>
</table>
The differences between the CLT and the solid model are now quite large even for the thinnest case, where they range from 32% to 49% and reach more than 400% for the thick case. Buckling modes differ significantly now. The same can be said of the differences between the FSDT and the 3D model. Finally, if shear correction factors are used, buckling loads are below the 3D model with differences between -4% and -9% for the thin case and around -36% to the thick case. In this latter case, the problem is that the predicted buckling modes are concentrated near the clamped area where the pre-critical bending deformation are larger and they do not resemble those obtained with the solid model.

5 CONCLUSIONS

This work studies the influence of shell models used in determining buckling loads in fully clamped compressed square plates and cylindrical panels subjected to lateral or axial load. Comparisons are made against 3D models. The cross sections considered are sandwich laminates and the shell models used are the classical laminate theory (CLT), the first order shear deformation order (FSDT) with and without the use of shear correction factors and the refined theory zigzag (RZT).

For square plates a combination of materials has been used in which the relationship between the mechanical properties of the core material and the outer layers is of the order of (1) $10^{-1}$ and (2) $10^{-3}$. The ratio between thickness and side are $1/100$ (thin) and $1/20$ (thick). It can be said that

- In case (1) the buckling loads and associated modes are very similar for the different models. Naturally they increase when more kinematic constraints are imposed but
For large thickness/side ratios the results with shell models may not be reliable because shell models can not capture local transverse effects.

In case (2) for thin plates, the buckling loads are of the same order for the different theories but FSDT and CLT give values of up to 30% in excess. For thicker plates buckling loads are one order of magnitude greater in the case of CLT and FSDT while the buckling modes are different from the 3D model.

The results provided by RZT and 3D models show an excellent agreement for both sections and thickness/side ratios.

The use of shear correction factors leads to good results for thin plates but not for thick plates where the buckling mode may be localized.

For cylindrical sectors only the combination of materials with properties ratio of an order of $10^{-3}$ has been considered. The sections analyzed have two ratios $t/D = 0.01$ (thin) and $t/D = 0.05$ (thick). Two load cases were alternatively considered, external pressure (hoop compression) and axial load (meridian compression).

The results obtained with RZT and 3D models are in excellent agreement for both load cases and thickness/radius ratios.

Both the CLT and the FSDT give buckling loads well above the 3D model especially for thick case, besides bifurcation modes are very different.
• The use of shear correction factors leads to non-conservative values of buckling loads and the critical modes differ significantly from the 3D model.

ACKNOWLEDGMENTS
The author gratefully acknowledge financial support from CONICET and SeCyT-UNC.

REFERENCES


INTERACTION BETWEEN SOIL AND BURIED PIPELINE

SARITA MARIA DE FIGUEIREDO SOUZA MELO\textsuperscript{1}, LÍCIA MOUTA DA COSTA\textsuperscript{2} AND IVALDO PONTES\textsuperscript{3}

\textsuperscript{1}Dept. of Mechanical Engineering, Federal University of Pernambuco, Brazil
Av. Acadêmico Hélio Ramos, s/n – 50740-530 Recife, Pernambuco, Brazil
E-mail address: sarita.figueredo@hotmail.com

\textsuperscript{2} Dept. of Civil Engineering, Federal University of Pernambuco, Brazil
Av. Acadêmico Hélio Ramos, s/n – 50740-530 Recife, Pernambuco, Brazil
E-mail address: licia@ufpe.br

\textsuperscript{3} Dept. of Civil Engineering, Federal University of Pernambuco, Brazil
Av. Acadêmico Hélio Ramos, s/n – 50740-530 Recife, Pernambuco, Brazil
E-mail address: ivaldo@ufpe.br

Key words: Buried pipelines, Finite Elements, Plasticity.

Abstract. The use of buried pipelines is increasingly common and plays an important role in oil industry. Its good structural performance is essential to oil or gas flow. From the engineering point of view the stress-displacement relation in buried pipelines is very important to ensure system integrity. The definition of the maximum load to be used in pipeline is very important in the design, since the displacement and rupture mechanism change significantly when changing the pipeline depth. In this paper the theory of plasticity will be used to study the mechanisms involved in the rupture of buried pipelines, instead of estimating the burden of interaction from the theory of bearing capacity of shallow foundations. It will also examine the possibility of lateral loads on piles. The results demonstrate the importance of lateral loads.

1 INTRODUCTION

Buried pipelines are under several loadings such as weight of soil, external pressure due to water depth, thermal stresses due to fluid temperature. Under such efforts, a pipe elongation is expected if the structure were free of supports. Indeed, the pipe movements are restricted due to soil-structure friction and anchors. Consequently, axial forces are developed along its length and the pipe behaves like a column. Vertical movement of the pipeline in soil is governed by mechanisms similar to the lateral loading of a circular pile up to its limit load [1]. Buckling may occur in order to relief the stresses and if it occurs in an uncontrolled way, a structure collapse can be extended over a long pipe length. However, according to Bruton et al. (2005) [2], inducing controlled lateral buckling may be a solution to relief compressive stresses and some methods to induce its formation are shown. The vertical and lateral movement occurrence depends on some variables such as pipe embedment, soil resistance and friction between soil-structure.
The prediction, control and simulation of these phenomena deal with plasticity theory and finite element method. The pipe is considered as a rigid body and the soil around it considered as a deformable body. The model simulates the soil mechanical response, discretized as 2-D finite elements under plane strain assumption. A velocity field is applied as a boundary condition at the interface pipe-soil in order to simulate the pipe movement, plastically deforming the soil mass.

The purpose of this paper is to evaluate, by numerical simulation, the necessary efforts that will plastically deform the soil mass around a shallow embedded pipe. The plastic behavior of the soil depends on the yield function and the soil parameters taken into account. Classical yield functions von Mises and Drucker-Prager are used in the analysis carried out.

2 KINEMATICS AND EQUILIBRIUM

Consider a body occupying an open bounded region $\mathcal{B}$ with a regular boundary $\Gamma$. Let $V$ denote the function space of all admissible displacement fields $u$ complying with the homogeneous boundary conditions prescribed on a part $\Gamma_u$ of $\Gamma$. The strain rate tensor fields $\varepsilon$ are elements of the function space $W$ and the tangent deformation linear operator $D$ maps $V$ onto $W$.

Let $W'$ be the space of stress fields $\sigma$. The internal work for any pair $\sigma \in W'$ and $\varepsilon \in W$ is defined by the duality product:

$$\langle \sigma, \varepsilon \rangle = \int_{\mathcal{B}} \sigma \cdot \varepsilon \ d\mathcal{B}$$

(1)

In the same way, $V'$ is the space of loads and the external power dissipated by a loads system $F \in V'$ on a displacement field $u \in V$ is given by the duality product:

$$\langle F, u \rangle = \int_{\mathcal{B}} b \cdot u \ d\mathcal{B} + \int_{\Gamma} \tau \cdot u \ d\Gamma$$

(2)

where $b$ and $\tau$ are body and surface loads respectively. Surface $\Gamma_\tau$ is the region of $\Gamma$ where tractions are prescribed ($\Gamma = \Gamma_u \cup \Gamma_\tau$ and $\Gamma_u \cap \Gamma_\tau$ is empty).

The equilibrium condition, relating a stress field and a load system prescribed $F \in V'$, is imposed by the principle of virtual work:

$$\langle T, Du \rangle = \langle F, u \rangle, \quad \forall u \in V$$

(3)

3 CONSTITUTIVE RELATIONS FOR ELASTIC IDEALLY PLASTIC MATERIALS

The stress field $\sigma$ in an elastic-ideally plastic body $\mathcal{B}$ is constrained to fulfill the plastic admissibility condition, i.e. it must belong to the set:

$$P = \{ \sigma \in W' \mid f(T \leq 0 \ in \ \mathcal{B}) \}$$

(4)

where $f$ is a $\mathbb{R}^n$ - vector valued function describing the yield criterion. The inequality above is then understood as constraining each component $f$, which is a regular convex function of $T$. 

54
to be nonpositive.

The plastic flow law relates plastic strain rates \( \dot{\varepsilon}^p \) to stresses \( \sigma \). This relation, which is not one-to-one, is considered next for the case of associative plastic flow.

Let us define the dissipation function as

\[
\chi(\dot{\varepsilon}^p) = \sup_{\sigma^* \in P} \left\langle T^*, \dot{\varepsilon}^p \right\rangle
\]

(5)

The stresses corresponding to a given plastic strain rate constitute the set of solutions of the above optimization problem, as stated by the principle of maximum dissipation [3,4,5]. This relationship complies with the second law of thermodynamics [5,6]. Using basic results of convex analysis we can state this constitutive relation in the following equivalent forms:

\[
\sigma \in \partial \chi(\dot{\varepsilon}^p) \iff \dot{\varepsilon}^p \in C_p(\sigma)
\]

(6)

where the subdifferential \( \partial \chi(\dot{\varepsilon}^p) \) is the set of all stress fields such that

\[
\chi(\dot{\varepsilon}^{p*}) - \chi(\dot{\varepsilon}^p) \geq \left\langle \sigma, \dot{\varepsilon}^{p*} - \dot{\varepsilon}^p \right\rangle, \quad \forall \dot{\varepsilon}^{p*} \in W
\]

(7)

and \( C_p(\sigma) \) is the cone of normals to \( P \) at \( \sigma \), which is only defined for \( \sigma \in P \). The elements of \( C_p(\sigma) \) are all the strain rate fields such that

\[
\left\langle \sigma^* - \sigma, \dot{\varepsilon}^p \right\rangle \leq 0, \quad \forall \sigma^* \in P
\]

(8)

It is worth mentioning the equivalence of equation (6) to the classical form of the plastic constitutive equations. In fact, when the set \( P \) is defined as in (4), i.e. by means of a function \( f \), it can be shown [7] that strain rates belonging to \( C_p(\sigma) \) are related to the stress, at any point of \( \mathcal{B} \), by the normality rule \( \dot{\varepsilon}^p = \nabla f(\sigma) \dot{\lambda} \). Here \( \nabla f(\sigma) \) denotes the gradient of \( f \) and \( \dot{\lambda} \) is the \( \dot{m} \)-vector field of plastic multipliers. At any point of \( \mathcal{B} \) the components of \( \dot{\lambda} \) are related to each plastic mode in \( f \) by the complementarity condition \( \dot{\lambda} \geq 0, f \leq 0 \) and \( f \cdot \dot{\lambda} = 0 \).

When the Drucker-Prager criterion is used to define the plastic admissibility conditions, the plastic function \( f \) is written as

\[
f(T) = 3\alpha \sigma_m + \frac{1}{\sqrt{2}} \|S\| - k
\]

(9)

where \( \alpha \) and \( k \) are material constants. For plane strain state \( k = c \cos \phi \), in which \( c \) is cohesion and \( \phi \) is the friction angle. When \( \phi = 0 \) Drucker-Prager criterion reduces to von Mises.
4 ANALYSIS CARRIED OUT

Some examples of buried pipelines subjected to vertical and lateral loading were analyzed. The pipeline is considered rigid and the soil as a deformable body. This doesn’t mean that the results are not valid for relatively flexible pipelines too, provided that the deformation of pipeline section is small compared to its dimensions, under the hypothesis of small strain (Kouretzis et al., 2014). It is also considered no friction between pipeline and soil. The examples carried out considered the pipeline buried at three different depths (H), the first one is at H=D/2, where D is the pipeline diameter. In the second example H=D and in the last one H=3D.

4.1 Pipeline buried at H=D/2

In this example it was used a finite element mesh with 1023 linear triangles and 551 nodal points (Figure 1).

![Finite element mesh for pipeline buried at H=D/2](image)

Figure 1: Finite element mesh for pipeline buried at H=D/2

The loading conditions applied for evaluating the vertical and lateral movement, considered prescribed displacement rate in pipeline as indicated in figure 2.

![Prescribed displacement rate for (a) vertical loading; (b) horizontal loading.](image)

Figure 2: Prescribed displacement rate for (a) vertical loading; (b) horizontal loading.
The soil was considered an elastic ideally plastic material with Von Mises criterion and associated flow law. The ultimate load of the incremental process was calculated by finite element simulation and it was obtained the ratio $F/cD = 9.24$. Figure 3 shows the distribution of the second deviatory stress, where it can be observed the higher values at the pipeline.

![Image](image1.png)

**Figure 3:** Deviatory stress distribution for vertical loading for Von Mises criterion.

The second example using the same mesh, considered the soil as an elastic ideally plastic material with Drucker Prager criterion material and associated flow law. It was considered a friction angle of $20^\circ$ for the soil. The ultimate load obtained by the incremental process calculated by finite element simulation gives the ratio $F/cD = 12.70$. The stress distribution was very similar to the first example, but achieving higher values.

![Image](image2.png)

**Figure 4:** Deviatory stress distribution for vertical loading for Drucker Prager criterion.
Another analysis was performed with the same data of the second example (elastic ideally plastic material with Drucker Prager material, associated flow law and friction angle of $20^\circ$ for the soil) but now applying horizontal displacement rate (Figure 2 (b)). In this case [8] the finite element simulator calculated the ultimate incremental load, at horizontal direction, according to the following ratio $F/cD = 5.72$.

The displacement achieved in this analysis can be observed in figures 5 (displacement vectors) and 6 (deformed mesh).

![Figure 5: Displacement vector due to horizontal loading](image5.png)

![Figure 6: Deformation due to horizontal loading](image6.png)
4.2 Pipeline buried at H=D

The finite element mesh used in this example has 1242 linear triangles and 665 nodal points (Figure 7).

![Finite element mesh for pipeline buried at H=D](image)

**Figure 7:** Finite element mesh for pipeline buried at H=D

As in the analysis for H=D/2, two examples were performed, both considered the soil an elastic ideally plastic material and associated flow law, but one with Von Mises criterion and another with Drucker Prager (friction angle of 20° for the soil). The ultimate load of the incremental process calculated by finite element simulation obtained the ratio F/cD = 10.38 and 13.82 respectively. The deviatory stress distribution (Figure 8) is very similar for both examples, but Drucker Prager obtained higher values than Von Mises.
4.3 Pipeline buried at H=3D

The same procedure for the previous analysis was used for pipeline buried at H=3D. For this example the finite element used is illustrated in figure 9, with 1496 linear triangles and 783 nodal points. Two examples were performed; both considered the soil an elastic ideally plastic material and associated flow law, but one with Von Mises criterion and another with Drucker Prager (friction angle of 20° for the soil). The ultimate load of the incremental process calculated by finite element simulation obtained the ratio F/cD = 10.42 e F/cD = 14.92 respectively. Again, the deviatory stress distribution (Figure 10) is very similar for both examples, but higher values was observed for Drucker Prager criterion.
**Figure 9:** Finite element mesh for pipeline buried at H=3D

**Figure 10:** Deviatory stress distribution for vertical loading for pipeline buried at H=3D
5 CONCLUSIONS

The increase of the buried depth increases the ultimate strength of the soil surrounding the pipe. However, as observed in these tests the growth of the ultimate load increases more slowly with increasing depth.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the support of this research by CNPq and PRH26.

REFERENCES


FINITE ELEMENT ANALYSIS OF MECHANICAL DEFORMATION PROCESSES IN SEDIMENTARY BASINS

ANDRÉ BRÜCH¹, SAMIR MAGHOUS²

¹Federal University of Rio Grande do Sul, Department of Civil Engineering
Av. Osvaldo Aranha 99, 3rd floor, Centro – CEP 90035-190 – Porto Alegre, RS, Brazil
andrebruch@hotmail.com

²Federal University of Rio Grande do Sul, Department of Civil Engineering
Av. Osvaldo Aranha 99, 3rd floor, Centro – CEP 90035-190 – Porto Alegre, RS, Brazil
samir.maghous@ufrgs.br

Key words: Sedimentary Basin, Finite Poroplasticity, Gravitational Compaction, Tectonics.

Abstract. The study of sedimentary basins is an important issue in the field of geophysics and geomechanics that seeks understanding the geological history and reconstructing the poromechanical history of many regions of the planet. The purpose of this work is to study the mechanical behavior of sedimentary basins through tridimensional numerical simulations of different fundamental processes involved in its formation, such as sediment deposition, gravitational compaction and tectonic-driven deformation. The process of mechanical compaction of a sedimentary basin in fully saturated conditions takes place through water expulsion from the porous material, thus resulting in grain repacking and volume reduction. In absence of tectonic-driven deformation, a deposited sedimentary layer compacts as the excess pore-pressure generated by its own weight dissipates progressively. Nevertheless, tectonic movements play an important role in the mechanical deformation of basins, accelerating or slowing down its compaction process or even leading to failure surfaces. One of the major difficulties involving this kind of simulation is connected with the occurrence of large change in porosity throughout the compaction process. The coupled nature of the deformation problem may be understood as follows: large strains modify the microstructure, which leads to a change in the poromechanical properties of the sediment material and thus affecting the basins response. This behavior requires that the poromechanical constitutive law be formulated in the framework of finite irreversible strains with account for hydro-mechanical and elasticity–plasticity couplings. The numerical simulations are performed through the finite element method with a shared memory multiprocessing interface. The geometry model consists in a prismatic rectangular mesh which is intended to represent an offshore sea bed where accumulation of a thick sedimentary layer occurs over tens of millions of years. After this long period of deposition and compaction of sediments under gravitational forces and excess pore-pressure dissipation, extensional or compressive tectonic movements are applied on the sides of the basin. Some important points of the theoretical and numerical models are discussed as well as a comparative analysis of the deformation process with and without tectonics.
1 INTRODUCTION

The study of sedimentary basins is an important issue in the field of geophysics and geomechanics that seeks understanding the geological history and reconstructing the poromechanical history of many regions of the planet. The potential applications include petroleum exploration, reserve assessment and production. They are important aquifers and a large variety of metallic and non-metallic minerals can be also found in this kind of geological formation. Detection and exploration of all these resources requires sophisticated prediction tools and technical methods, which have been developed in many ways along the past decades. Much of this evolution is due to the oil industry, where a need for knowledge on the subject of oil reservoirs has led to enormous expansion of research on modern sedimentary environments [1].

In this context, several numerical models specifically devised for simulating formation and evolution of sedimentary basins have been developed. These models are used to simulate natural processes such as sediment erosion, deposition, compaction, deformation induced by tectonic movements, heat transport, and the formation, migration and accumulation of hydrocarbons, which occur during the evolution of sedimentary basins. The main objective of appropriate modeling is the reduction in exploration costs and risks [2].

However, the progress in this field has been hampered by the absence of comprehensible models to describe the rheological behavior of the constituent material of the problem. Since it is a multidisciplinary problem, incorporating all the coupled phenomena controlling the basin formation within the computational model turns to be a very difficult task [3].

Sedimentary basins result from deposition of an appreciable amount of sediments originated from destruction of any type of rock which are transported and deposited in different types of environments and transformed into rock through natural phenomena involving mechanical and chemical processes. Their occurrence in diverse geological settings is usually associated with plate tectonic activity. The process of mechanical compaction of a sedimentary basin in fully saturated conditions takes place through water expulsion from the porous material, thus resulting in grain repacking and volume reduction. Chemical compaction involves mineral dissolution and precipitation, and is a function of mineral stability and kinetics of precipitation in sediments. Mechanical phenomena prevail in the upper layers of sedimentary basins, whereas chemical compaction dominates for the deeper layers as stress and temperature increase [4].

Disregarding the chemical aspects, the purpose of this work is to study the mechanical behavior of sedimentary basins. Different fundamental processes involved in its formation, such as sediment deposition, gravitational compaction and tectonic-driven deformation are simulated through tridimensional numerical simulations via finite element method with a shared memory multiprocessing interface. The simulations are performed on a prismatic rectangular mesh which is intended to represent an offshore sea bed where accumulation of a thick sedimentary layer occurs over tens of millions of years. After this long period of deposition and compaction of sediments under gravitational forces and excess pore-pressure dissipation, extensional or compressive tectonic movements are applied on the sides of the basin. Some important points of the theoretical and numerical models are discussed as well as a comparative analysis of the compaction process with and without tectonics.
2 MATHEMATICAL MODEL

At the macroscopic scale, a porous medium can be viewed as the superposition of a solid continuum related to the deformable skeleton and a fluid phase occupying the porous space. In order to define the poromechanical problem under isothermal and quasi-static conditions, two governing equations referring to momentum and mass balance must be specified. By considering the balance of linear momentum, the equilibrium equation for the porous continuum is:

\[ \text{div}\sigma + \rho g = 0 \]  

(1)

where \( \sigma \) is the Cauchy total stress tensor, \( g \) is the acceleration of gravity and \( \rho = \rho^f (1 - \phi) + \rho^s \phi \) is the density of the porous material, which can be computed from the Eulerian porosity \( \phi \) (i.e. the pore volume fraction in the current configuration), the fluid density \( \rho^f \) and the solid grains density \( \rho^s \).

Denoting by \( \phi = J \phi \) the Lagrangian porosity, where \( J \) is the jacobian of the geometrical transformation of the skeleton particle, the second field equation refers to the fluid mass balance:

\[ \frac{d}{dt}(\rho^f \phi) + J \text{div}(\rho^f q) = 0 \]  

(2)

where \( q \) is the filtration vector defined according to the Darcy’s law:

\[ q = k \left( -\nabla p + \rho^f g \right) \]  

(3)

where \( k \) is the permeability tensor, which is reduced to \( k = k_1 \) when isotropy is taken into account, and \( p \) is the pore-pressure.

In the range of infinitesimal strains, the linear poroelastic state equations are given according to Biot’s theory. They relate the stress and pore-pressure to the elastic strain \( \varepsilon^e = \varepsilon^e - Bp \) and reversible porosity change \( \phi^r 

\[ \sigma = C : \varepsilon^e - Bp \]  

(4)

\[ p = M \left( \phi^r - B : \varepsilon^e \right) \]  

(5)

where \( C \) is the fourth order tensor of drained elastic moduli, \( B \) is the Biot tensor and \( M \) is the Biot modulus. In what follows, elastic isotropy will be assumed for the porous material. Thus, \( B = b I \)

(6)

\[ C = (K - 2\mu/3)I + 2\mu I \]  

(7)

where \( b, K \) and \( \mu \) denote respectively the Biot coefficient, the bulk and shear moduli.
The macroscopic equations of state of a fully saturated poroelastoplastic material generally assume a total independence of the elastic properties with respect to the plastic behavior. However, in the context of large strains, such an assumption is clearly questionable as the large macroscopic plastic strains are associated with an irreversible evolution of the microstructure, which is responsible for variations of the macroscopic elastic properties of the porous medium. We shall refer to this phenomenon as elasticity–plasticity coupling [5].

In the context of finite poroplasticity and considering the evolution of the poromechanical properties, the macroscopic rate equations of state take the following rate form [5]:

\[
\frac{D_t \sigma^e}{Dt} = \sigma^e + \sigma^e \Omega - \Omega \sigma^e = \dot{C} : C^{-1} : \sigma^e + \ddot{C} : (d - d^p)
\]  

(8)

\[
\dot{p} = M \left( \dot{\phi} - \phi^p \right) - btr \left( C^{-1} : \sigma^e \right) - btr \left( d - d^p \right) + \frac{M}{M} p
\]  

(9)

where \( \sigma^e = \sigma + B \rho \) is the Biot effective stress, \( \Omega \) is the spin tensor, \( d \) is the total strain rate and \( d^p \) is the plastic strain rate.

In the framework of infinitesimal elastic strains, it is possible to assign the following approximation to the plastic part of the geometrical transformation of the skeleton particle \( J^p = J \). Based on this assumption together with the assumption of plastic incompressibility for the solid phase, it can be readily shown [3] that:

\[
\varphi = 1 - \frac{1 - \phi_0}{J^p}
\]  

(10)

The complementary relationships that define the constitutive model are introduced as follows. The yield surface of the porous material is that of the modified Cam-Clay model [6]:

\[
f \left( \sigma^e, p_c \right) = \frac{3}{2} s \cdot s + M_{cs} p \left( p' + p_c \right)
\]  

(12)

where \( p_c \) is the consolidation pressure and represents the hardening parameter of the model. The constant \( M_{cs} \) is the slope of the critical state line. The plastic criterion depends on the Terzaghi effective stress \( \sigma^e = \sigma + p_1 \) through

\[
s = \sigma - \frac{1}{3} tr ( \sigma )
\]

and

\[
p' = \frac{1}{3} tr ( \sigma' )
\]  

(11)

The hardening law has been the object of special concern. Based on micromechanical reasoning, a specific formulation has been developed for the evolution of \( p_c \) [7]:

\[
p_c \left( J^p \right) = \frac{p_{c0}}{\ln \left( \phi_0 \right)} \ln \left( 1 - \frac{1 - \phi_0}{J^p} \right)
\]  

(13)

The main advantage of this hardening law with respect to classical ones lies in the fact that it avoids the development of negative porosities under high isotropic compression [3,8].
It is worth noting that neither the dependence of bulk and shear elastic moduli with respect to mean effective stress considered in the original Cam-Clay model nor the associated hardening rule [9] are used in the present modeling. Actually, a linear elastic behavior is assumed herein, and only the shape of plastic yield and associated plastic flow rule are borrowed from Cam-Clay model.

Since the porous medium deformation is characterized by significant changes in porosity, the evolution of the elastic stiffness with large irreversible strains shall be considered. The idea is to resort to micromechanical estimates to capture the influence of the plastic strains on the poroelastic properties. For sake of simplicity, the anisotropy induced during the loading process is disregarded. Accordingly, the bulk \( K \) and shear \( \mu \) moduli of the porous medium now appear as functions of the Eulerian porosity \( \phi \) as well as of the elastic properties of the solid phase \( k^s \) and \( \mu^s \). More precisely, we herein adopt the Hashin–Shtrikman upper bounds which are known to reasonably model the elastic properties of isotropic porous media [10].

\[
K(\phi) = \frac{4k^s\mu^s (1-\phi)}{3k^s\phi + 4\mu^s} \quad \mu(\phi) = \frac{\mu^s (1-\phi)(9k^s + 8\mu^s)}{k^s (9 + 6\phi) + \mu^s (8 + 12\phi)}
\] (14)

Biot coefficient \( b \) and Biot modulus \( M \) are deduced from the following equations [11]:

\[
b(\phi) = 1 - \frac{K(\phi)}{k^s} \quad M(\phi) = \frac{k^s}{b(\phi) - \phi}
\] (15)

The effects of microstructural changes on the evolution of the permeability coefficient of the porous medium \( k \) may be modeled by means of the Kozeny–Carman formula:

\[
k(\phi) = k_0 \frac{\phi^3 (1-\phi_0)^2}{\phi_0^3 (1-\phi)^2}
\] (16)

3 FINITE ELEMENT DISCRETIZATION

Assessment of the poromechanical state requires the determination of the temporal evolution of the geometric transformation as well as the pore-pressure changes. This shall be achieved by solving the boundary value problem defined by the set of governing equations together with the constitutive and complementary equations. The particularity of large strains is that all equations actually refer to the mechanical system in its current configuration, which is a priori unknown. The finite element procedure used for assessing the evolution of the porous medium under consideration will be outlined hereafter [3,5].

The analysis is based on the implementation of the updated Lagrangian scheme [12]. This approach is based on the same procedures used by total Lagrangian formulations, but instead of being referred to the initial configuration, all static and kinematic variables are referred to the last calculated configuration, say at time \( t \). The unknown variables are then updated in each step time \( \Delta t \). The displacement \( U \) of the skeleton particles between \( t \) and \( t + \Delta t \) is defined as:
\[ U = \hat{x}^{t+\Delta t} - \hat{x}' \]  

(17)

where \( \hat{x}' \) (resp. \( \hat{x}^{t+\Delta t} \)) denotes the particle coordinate at time \( t \) (resp. \( t + \Delta t \)). Pore-pressure difference at points similar within the skeleton transformation between \( t \) and \( t + \Delta t \) is:

\[ P = p(\hat{x}^{t+\Delta t}) - p(\hat{x}') \]  

(18)

The discretized form of the problem is obtained from weak formulation of the equilibrium and fluid mass balance equations at time \( t + \Delta t \). Twenty-nodes hexahedra are used for geometry discretization (Figure 1). A piecewise quadratic polynomial function is adopted to approximate the displacement, while piecewise linear function is adopted for pore-pressure variations. The resulting system from finite element discretization is given below:

\[
\begin{bmatrix}
K_{UU} & K_{UP} \\
K_{PU} & K_{PP}
\end{bmatrix}
\begin{bmatrix}
U \\
F
\end{bmatrix}
= 
\begin{bmatrix}
F_U(U, P) \\
F_P(U, P)
\end{bmatrix}
\]

(19)

where \( K_{UU} \) are the global stiffness sub-matrices and \( F_J \) the global force sub-vectors. Global vectors \( U \) and \( P \) are respectively the nodal displacements and pore-pressure difference. For a single element, \( K_{UU} \) is a 60×60 matrix, \( K_{UP} \) 60×8, \( K_{PU} \) 8×60 and \( K_{PP} \) 8×8. On the other hand, \( F_U \) is 60×1 and \( F_P \) is 8×1.

It is emphasized that, for a given configuration at time \( t \), the system above is highly non-linear due to the physical non-linearities (plasticity) and geometrical non-linearities (large strains). An iterative method is adopted for solving this system by implementing an appropriate algorithm until it is satisfied up to a required tolerance.

A main specificity of the problem lies in the fact that a sedimentary basin is an open system due to the continuous accretion of material at the top of the basin during the deposition phase. This requires an appropriate technique to overcome the difficulty of dealing with an open system in the context of the finite element method. The activation/deactivation method is used to simulate the accretion process [3]. In the framework of such method, the real open material system is simulated as a fictitious closed one.

Figure 1 illustrates the basin construction problem, where \( L_0 \) corresponds to the sea level and \( H(t) \) corresponds to the top of the sediment layer. The evolution in time of the latter must of course correspond to that of the real system. The simulation of the accretion phase takes place by subdividing its period \( T \) into \( n \) subintervals \([t_{i-1}, t_i]\) with \( t_0 = 0 \) and \( t_n = T \). During the time increment \( \Delta t_i = t_i - t_{i-1} \), the sediments supply corresponds to a height of \( \Delta H_i \)

\[ \Delta H_i = \int_{t_{i-1}}^{t_i} \frac{M_d(x,y,t)}{\rho_0(x,y,t)} dt \]  

(20)

where \( M_d \) is the rate of sediment mass accretion per unit area and \( \rho_0 \) represents the mass density of the deposited material.
Before sediment deposition, all finite element layers constituting the system behave as seawater. This is achieved by setting a hydrostatic total stress and pore-pressure state to these elements. It is also necessary to set the sea water density \( \rho^f \) and a high permeability coefficient (when compared to the sediment material) \( |k| >> k_0 \). In these elements, stress and pore-pressure are set to:

\[
\sigma = -\rho^f g (L_0 - H_i) \\
p = \rho^f g (L_0 - H_i)
\]

(21)

For every advanced subinterval \( \Delta t_i \), the properties of a corresponding layer with thickness \( \Delta H_i \) are changed in conjunction with the hydraulic and elastic properties of the deposited material. This process starts at the bottom of the basin \( (z=0) \) and continues upwards until the accretion phase ends.

**Figure 1**: Schematic geometry of the basin and the finite element characteristics

### 4 COMPUTATIONAL RESULTS

The initial geometry of the basin consists in a rectangular prism with 6km height and 12km×12km base. The mesh is divided into 20 equal parts in each direction, resulting in 8000 hexahedra elements and 35721 nodes. The considered situation has the following characteristics [13]: the sea bed located at \( z=0 \) is considered rigid and impermeable and the contact with the basin is frictionless; basin sides are impermeable; the sea level is considered fixed at \( L = 8000 \) m; the rate of sediment deposition is constant and equal to 0.1 mm/year along the accretion phase of the basin, which starts at \( t = 0 \) and takes \( T_{sed} = 60 \text{ million years} \), resulting in a quantity of material which would correspond to a vertical column of 6km in the absence of compaction, that is if the deposited material was rigid.
Same initial properties are considered for the sediment material along the accretion phase: density $\rho_0 = 1.37 \times 10^3 \text{ kg/m}^3$, porosity $\phi_0 = 0.72$, Young modulus $E_0 = 10^3 \text{ MPa}$, Poisson’s ratio $v_0 = 0.33$, Biot coefficient $k_0 = 0.98$, Biot modulus $M_0 = 2.22 \times 10^5 \text{ MPa}$, coefficient slope of the critical state line $M_c = 1.2$, permeability $k_0 = 10^{-9} \text{ m}^2 \times (\text{MPa} \times \text{s})^{-1}$, consolidation pressure $p_{c0} = 1.5 \text{ MPa}$, sea water density $\rho^f = 10^3 \text{ kg/m}^3$.

4.1 Gravitational compaction

Figure 2 sketches the basin configuration and the boundary conditions of the model for any section in the $xz$ plane. From deformation viewpoint, it consists in a one-dimensional problem as the basin is kept under oedometric conditions. Thus, gravitational compaction process could be modeled by a simple column of elements. We justify the use of a three-dimensional mesh because the result of this analysis is used subsequently for tectonic simulations.

![Figure 2: Basin configuration and boundary conditions of the model](image)

![Figure 3: Compaction law of the basin](image)
Figure 3 illustrates the compaction law of the basin. As it can be seen, the thickness of the basin is about $H = 3640 \text{ m}$ at the end of accretion phase $t = T_{\text{sed}} = 60 \text{ My}$, which represents a compaction level of 40%. At time $t = 10T_{\text{sed}} = 600 \text{ My}$, the thickness of the basin is almost stabilized at $H = 2548 \text{ m}$, corresponding to 57.5% compaction level.

Since it is a one-dimensional problem, the description of the poromechanical properties of the basin may be performed by a vertical profile of any chosen position on its horizontal plane. The coordinate $(x=0; y=0)$ was chosen to this end. Figure 4 presents the pore-pressure and porosity profiles of three different ages from compaction law: 1) end of accretion phase, $t = 60 \text{ My}$; 2) configuration reached at $t = 600 \text{ My}$; 3) an intermediate configuration with $H = 3000 \text{ m}$ at $t = 80 \text{ My}$. The hydrostatic profile is also illustrated. It can be seen that although case $t = 600 \text{ My}$ is nearly asymptotic in relation to its compaction law, its pore-pressure profile is still far from hydrostatic. The Kozeny-Carman formula used to quantify the evolution of the permeability coefficient indicates that $k \to 0$ when $\phi \to 0$. This explains the long time required for total dissipation of excess pore-pressure in the lower layers of the basin, i.e. the time required to attain the asymptotic state. Regarding the porosity profile, an important remark lies in the fact that the upper layers remained in the elastic domain, since every porosity change is associated to volumetric plastic strains of the porous material.

![Figure 4: Pore-pressure and porosity profiles in purely gravitational compaction](image_url)
4.2 Tectonic sequences

Tectonics phase is simulated by imposing on the sides of the basin a prescribed displacement $\delta u$ for each time increment $\delta t$. Therefore, the velocity of tectonic plates can be adjusted by simply changing any of these values, since $V = \delta u / \delta t$. This can be done separately for each horizontal direction $x$ or $y$. The boundary conditions applied to the simulation of tectonics phase for any section in the $xz$ plane are shown in figure 5.

Figure 5: Boundary conditions for tectonic sequences

It was adopted as initial state of the basin the configuration reached at $t = 80$ My during gravitational compaction simulation. For a first approach on the modeling of tectonics, the idea is to analyze the influence of imposed velocities on basin behavior, so the displacements were applied until only 1% of lateral deformation was reached for each case. The tests were performed by applying the same prescribed velocity in both horizontal directions of the basin $V_x = V_y$. Three velocities were used for extensional and compressive tectonic motion: $\pm 1 \text{mm/ year}$, $\pm 0.1 \text{mm/ year}$ and $\pm 0.01 \text{mm/ year}$.

The reason why the modified Cam-Clay model was chosen as plastic model is the comprehensive manner it describes the plastic behavior of the porous material. One interesting analysis concerns the ratio between the deviatoric stress $q = \sqrt{(3/2)s : s}$ and the mean effective stress $p'$. The obtained result for $-q/p'$ indicates the material behavior when compared to the slope of the critical state line $M_{cs}$. If $-q/p' < M_{cs}$ the material is in compaction process. If $-q/p' > M_{cs}$ it is in dilation process, and if $-q/p' = M_{cs}$ it means that the porous material lies in undrained conditions.

Figure 6 presents the $-q/p'$ profiles for extensional and compressive tectonic simulations after 1% of imposed lateral deformation for each case of applied velocity (load). In view of comparison, the $-q/p'$ relation obtained for $t = 80$ My before tectonic motion is presented as well. Before analyzing the obtained results, it is important to note that the ratio between the deviatoric and the mean effective stress has no physical meaning in the upper elastic layer, as this relation regards to the plastic behavior of the material. It is recalled that the adopted slope for the critical state line of the sediment material is $M_{cs} = 1.2$. 
As expected, the \( t = 80 \text{ My} \) profile indicates the basin was in contractive state before tectonic motion. Nevertheless, it can be seen that different plate velocity implies different basin behavior. The highest applied velocity \( \pm 1 \text{ mm/year} \) has led the basin to undrained behavior throughout all its thickness for extensional tectonics, and up to \( Z = 1500 \text{ m} \) for the compressive case. As the applied velocity decreases, basins layers undergo compaction. This change of behavior takes place from the upper to the lower layers of the basin.

A quantity that characterizes the macroscopic evolution of the basin is defined by the compaction rate \( V_c = \frac{dH}{dt} \) obtained from the compaction law presented in Figure 3. The speed of this phenomenon is associated with the basin capacity to dissipate excess pore-pressure by draining the fluid contained within its pores. The compaction law indicates that \( V_c = 0.015 \text{ mm/year} \) for \( t = 80 \text{ My} \). Thus, the imposition of loads at velocities exceeding this value should take the basin to undrained behavior, since the rate of increase of excess pore-pressure generated by these loads exceeds the dissipative capacity of the basin.

Although the present analysis of imposed lateral movements has been made considering a simplified scenario, the results have given evidence that tectonics can lead the basin to undrained behavior and stress points may migrate from contractive to dilation state. This numerical tool may provide useful guidelines in the perspective to predict occurrence of failure surfaces (faults) induced by extensional and compressive tectonics movements.
5 CONCLUSIONS

- Starting from the constitutive and numerical model briefly presented, three important phases involved in the mechanical processes of a sedimentary basin formation were simulated: the sediments accretion phase, its compaction due to gravitational forces and pore-pressure dissipation, and deformations imposed by tectonics.

- At the current stage of this research, chemical compaction due to intergranular pressure-solution (IPS) is being added to the constitutive and numerical model since it represents a major deformation mechanism of deep oceanic basins. Deeper investigations of tectonics deformation process are also being held.

REFERENCES

A NON-ISOTHERMAL CONSOLIDATION MODEL FOR GRADIENT-BASED POROPLASTICITY

JAVIER L. MROGINSKI\textsuperscript{1,4}, GUILLERMO ETSE\textsuperscript{2,4}, MARIANELA RIPANI\textsuperscript{3}

\textsuperscript{1} Applied Mechanical Dept., Universidad Nacional del Nordeste, Las Heras 727, Resistencia, Chaco, Argentina. javierm@ing.unne.edu.ar

\textsuperscript{2} Center for Numerical and Computational Methods in Engineering, National University of Tucum\textsuperscript{n}n, Argentina Maipu 780, 4000 Tucum\textsuperscript{n}, Argentina. getse@herrera.unt.edu.ar

\textsuperscript{3} Materials and Structures Laboratory, University of Buenos Aires, Argentina mripani@fi.uba.ar

\textsuperscript{4} National Scientific and Technical Research Council (CONICET)

Key words: Porous media, Gradient Theory, Thermodynamic Consistency, Non-isothermal

Abstract. In this work, the thermodynamically consistent non-local model for concretes subjected to high temperatures originally proposed by Ripani, et al (2014) \cite{1} is extended, in order to evaluate the failure behavior of partially saturated quasi brittle materials like soils. This new formulation follows the gradient-based poroplastic theory proposed by the authors \cite{2}, moreover, introduces the temperature as an additional variable of the internal characteristic length. The non-local effect is achieved assuming that the internal variables are the only ones of non-local character. Hence, both \(q\) and \(q_\alpha\) will be considered as arguments in the free energy. On the other hand, in order to reproduce the softening behavior of partially saturated soils the saturation degree as well as the confinement level should be considered in the internal characteristic length calibration. Finally, the FE implementation of this enriched constitutive theory with selective C1 and C0 interpolation functions for the internal variables and the kinematic fields, respectively, is discussed.

1 INTRODUCTION

The mechanics of porous media constitutes a discipline of great relevance in several knowledge areas like Geophysics, Biomechanics and Materials Science. Its main aim is
the description of the kinematic and pore pressure of porous continua when subjected to arbitrary mechanical and/or physical actions.

The definitive advantages of porous mechanics to macroscopically describe or predict complex response behavior of cohesive-frictional materials based on fundamental aspects of their microstructure while accounting for the hydraulic properties and their influence in the resulting failure mechanism were recognized by several authors in the scientific community [3, 9]. The influence of the non-isothermal condition in the failure behaviour was also evaluate by [22, 17]. Consequently, a tendency to replace the theoretical framework of classical continuum mechanics with that of non-linear porous mechanics was observed. Firstly this task took place in case of soil mechanics, see a.o. [7, 5], and subsequently in the field of concrete, see a.o. [27, 19] and, furthermore, of biomaterials , see a.o. [16, 20].

Further development in classical continuum mechanics was the extension to non-local concepts. The main aim was the regularization of post peak response behavior regarding mesh size and element orientation in case of finite element analyses, based on fundamental aspects of the material microstructure, see a.o. [25, 1, 32].

In recent years significant progresses and relevant contributions were made in non-local gradient formulations for non-porous materials. Thermodynamic frameworks were considered in the proposals of [1, 21, 28, 8, 32]. Considerations of material anisotropy in the formulation of internal variables evolution laws in case of gradient plasticity are due to [2, 29]. Formulation of gradient enhanced coupled damage-plasticity material models and related finite element implementations, see [25, 12, 6].

Recently, non-local concepts were extended for the formulation of porous material models, see a.o. [11, 10, 15, 14]. In spite of the strong development of constitutive modelling for porous media, explained before, there is still a need of thermodynamically consistent theoretical frameworks. This is particularly the case of non-local models for porous materials. Thermodynamic concepts should lead to dissipative stress formulations in hardening and softening regimes that allow non-constant descriptions of the internal variables of non-local character to accurately predict the sensitivity of porous material failure behavior to both confinement and saturation levels.

In this work, the thermodynamically consistent formulation for non-porous gradient-based elastoplasticity by Vrech and Etse [32] which follows general thermodynamic approach proposed by Svedberg and Runesson [25] for non-local damage formulation is extended for porous media. Main feature of present proposal is the definition of a gradient-based characteristic length in terms of both the governing stress and hydraulic conditions to capture the variation of the transition from brittle to ductile failure mode of cohesive-frictional porous materials with the confinement level and saturation [14].

2 Thermodynamics of Porous Continua

The postulate of local state stipulates that the internal energy of a homogeneous system is independent of the evolution rate and it can be characterized by the same state variables as the ones characterizing equilibrium states. The postulate of local state is extended to
porous continua by considering that their thermodynamics are obtained by adding the thermodynamic contributions of each constituent, that are the solid skeleton and the fluid continuum.

2.1 First law of the Thermodynamics

The first law of Thermodynamics expresses the conservation of energy in all forms. The energy $\dot{E}$ of a system can be expressed as the sum of its kinetic $\dot{K}$ and internal energy $\dot{E}$ of each component of this system. Considering a body occupying the volume $\Omega$, with boundary $\partial\Omega$, the first law of the thermodynamic is expressed as

$$\dot{E} = \dot{K} + \dot{E} = P_{\text{ext}} + Q$$

with

$$\dot{E} = \frac{d}{dt} \int_{\Omega} e \, d\Omega ; \quad \dot{K} = \frac{1}{2} \int_{\Omega} \rho_s (1 - \phi) |\dot{u}_i \dot{u}_i| + \rho_f |w_i w_i| \, d\Omega$$

$$Q = \int_{\Omega} \rho r \, d\Omega - \int_{\partial\Omega} h_i n_i \, d\partial\Omega ; \quad P_{\text{ext}} = \int_{\Omega} \rho b_i \dot{u}_i \, d\Omega + \int_{\partial\Omega} \sigma_{ij} n_i \dot{u}_j - \frac{p}{\rho_f} w_i n_i \, d\partial\Omega$$

Here, internal energy density (per unit mass) $e$ is the internal energy density (per unit mass), body force $b_i$ is the body force, total stress $\sigma_{ij}$ is the stress, pore pressure $p$ is the pore pressure, heat source density $r$ is a heat source density and heat flux $h_i$ is the heat flux. The displacement displacement $u_i$, the unit normal vector on $\partial\Omega$, $n_i$, and the mass density $\rho$, were also included.

Considering the equilibrium equation, the explicit form of the internal energy density for local dissipative porous material follows from Eq.(1) as

$$\rho \dot{e} = \sigma_{ij} \dot{e}_{ij} - h_f M_{i,j} - h_{i,i} + \rho r$$

being $\phi$ the porosity, $h_f$ the fluid specific enthalpy and $M_i$ the fluid flow vector.

2.2 The Second law of Thermodynamics

While, the first law states the conservation of energy in all of its forms, the second law states that the energy can only deteriorate. The second law introduces a new physical quantity, the entropy, which can only increase when an isolate system is considered. Let entropy volume density (per unit mass) $s$ be an entropy volume density (per unit mass), and according to the second law of thermodynamic the entropy $S$ of a thermodynamic system can not decrease. Thus

$$\dot{S} + Q_\theta \geq 0$$

with
\[ \dot{S} = \int_{\Omega} \rho \dot{s} + (sfM_i)_i \, d\Omega \quad ; \quad \mathcal{Q}_\theta = \int_{\Omega} \frac{\rho r}{T} \, d\Omega - \int_{\partial\Omega} \frac{n_i h_i^i}{T} \, d\partial\Omega \] (5)

being \( \mathcal{Q}_\theta \) the entropy flux and \( T \) the temperature. Transforming the surface integral of Eq. (5) into a volume integral, it follows that the volume integral in Eq. (4) must be non-negative for any system \( \Omega \), which yields

\[ \rho \dot{s} + (sfM_i)_i + \left( \frac{h_i^i}{T} \right)_i - \frac{\rho r}{T} \geq 0 \] (6)

Then, invoking the first law by Eq. (3) and introducing the Helmholtz’s free energy \( \Psi = e - sT \) as well as the free enthalpy of the fluid per mass unit (or Gibbs potential) \( g_f = h_f - sfT \), the following expression is attained

\[ \sigma_{ij} \dot{\varepsilon}_{ij} - g_f M_i^i - s \dot{T} - \dot{\Psi} - M_i (sfT_{i;i} + g_{f,i}) - \frac{h_i^i}{T} T_{i;i} \geq 0 \] (7)

Finally, considering the mass balance equation, \( \dot{\mathbf{m}} + M_{i;i} = 0 \) the Eq. (7) can be rewritten in the form

\[ \Phi_s + \Phi_f + \Phi_T \geq 0 \] (8)

with

\[ \Phi_s = \sigma_{ij} \dot{\varepsilon}_{ij} + g_f \dot{\mathbf{m}} - s \dot{T} - \dot{\Psi} \] (9)

\[ \Phi_f = -M_i (sfT_{i;i} + g_{f,i}) \] (10)

\[ \Phi_T = -\frac{h_i^i}{T} T_{i;i} \] (11)

The first component of Eq. (8) is related to the skeleton dissipation, \( \Phi_s \), the second accounts for the viscous dissipation due to the relative motion of the fluid with respect to the skeleton and the last source of dissipation, \( \Phi_T \), involves the temperature gradient \( T_{i;i} \) and therefore is related to the dissipation due to heat conduction.

Owing to the additive character of the Helmholtz free energy and entropy, \( \Psi = \Psi_s + m\Psi_f \) and \( s = s_s + m sf \), the fluid state equations and the relation \( m = \rho_f \phi \) allow to express \( \Phi_s \) as

\[ \Phi_s = \sigma_{ij} \dot{\varepsilon}_{ij} + p \dot{\phi} - s_s \dot{T} - \dot{\Psi}_s \] (12)

This expression of the skeleton dissipation \( \Phi_s \) matches the standard expression of the dissipation of a solid phase. Indeed, the strain work rate of an ordinary solid would reduce to the term \( \sigma_{ij} \dot{\varepsilon}_{ij} \). In the case of a porous continuous, the strain work rate related to the skeleton is obtained by adding \( p \dot{\phi} \), to account for the action of the pore pressure on the skeleton through the internal walls of the porous network.
3 Thermodynamically consistent gradient-based thermo-poro-plastic theory

The thermodynamic framework of classical or local plasticity is extended to non-local gradient-based elastoplastic porous material subjected to high temperatures. Following [23, 4] we assume that arbitrary thermodynamic states of the dissipative material during isothermal processes are completely determined by the elastic strain \( \varepsilon_{ij} = \varepsilon_{ij}^e - \varepsilon_{ij}^p \) or the temperature \( T \), the elastic entropy \( s^e = s - s^p \) and the internal variables \( q_\alpha \) with \( \alpha = s, p \) for solid or porous phase, which are considered here as scalar variables.

When considering poroplastic materials the elastic porosity \( \phi^e = \phi - \phi^p \) needs also to be included as a thermodynamic argument, see [4]. Based on [25, 32, 14] we further assume that the internal variables \( q_\alpha \) are the only ones of non-local character. Hence, both \( q_\alpha \) and \( q_{\alpha,i} \) will appear as arguments in the free energy \( \Psi_s \), such that

\[
\Psi_s = \Psi_s (\varepsilon_{ij}^e, \phi^e, T, q_\alpha, q_{\alpha,i})
\]  

Upon differentiation of Eq. (13) and combining with the intrinsic dissipation of Eq. (7) on the whole domain \( \Omega \), integrating the gradient term by parts and applying the Divergence Theorem, it follows

\[
\int_\Omega \left[ (\sigma_{ij} - \frac{\partial \Psi_s}{\partial \varepsilon_{ij}^e}) \dot{\varepsilon}_{ij} + (p - \frac{\partial \Psi_s}{\partial \phi^e}) \dot{\phi} + \left( -s_s - \frac{\partial \Psi_s}{\partial T} \right) \dot{T} + \frac{\partial \Psi_s}{\partial \varepsilon_{ij}^p} \dot{\varepsilon}_{ij}^p + \frac{\partial \Psi_s}{\partial \phi^p} \dot{\phi}^p \right]
\]

\[
+ \sum_\alpha Q_\alpha \dot{\alpha} \right] \, d\Omega + \int_{\partial \Omega} \sum_\alpha Q_\alpha^{(b)} \dot{\alpha} \, d\partial \Omega \geq 0 (14)
\]

The stress dissipation on the boundary \( \partial \Omega \) is defined as \( Q_\alpha^{(b)} = -\frac{\partial \Psi_s}{\partial q_{\alpha,i}} n_i \), also, in the domain \( \Omega \), the dissipative stress \( Q_\alpha \) can be decomposed into the local and non-local components, local dissipative stress \( Q^{loc}_\alpha \) and non-local dissipative stress \( Q^{nloc}_\alpha \), respectively

\[
Q_\alpha = Q^{loc}_\alpha + Q^{nloc}_\alpha
\]

with

\[
Q^{loc}_\alpha = -\frac{\partial \Psi_s}{\partial q_\alpha} ; \quad Q^{nloc}_\alpha = -\left( \frac{\partial \Psi_s}{\partial q_{\alpha,i}} \right) , i
\]

In standard form (as for local theory), it is postulated that the last inequality must hold for any choice of domain \( \Omega \) and for any independent thermodynamic process. As a result, Coleman’s equation are formally obtained like for the local continuum theory.

\[
\sigma_{ij} = \frac{\partial \Psi_s}{\partial \varepsilon_{ij}^e} ; \quad p = \frac{\partial \Psi_s}{\partial \phi^e} ; \quad s_s = -\frac{\partial \Psi_s}{\partial T}
\]

being the dissipative energy
\[
\mathcal{D} = \sigma_{ij} \dot{\varepsilon}_{ij}^p + p \dot{\phi}^p + \sum_{\alpha} Q_{\alpha} \dot{q}_{\alpha} \geq 0 \quad \text{in } \Omega \tag{18}
\]

\[
\mathcal{D}^{(b)} = \sum_{\alpha} Q_{\alpha}^{(b)} \dot{q}_{\alpha} \geq 0 \quad \text{on } \partial \Omega \tag{19}
\]

In the particular case of non-porous material \((p = 0)\) above equations take similar forms to those obtained by [25, 32] for isothermal situations.

From the above Eq. (18) and Eq. (19), it can be concluded that the difference between this simplified non-local theory and the local one is the additional gradient term in the expression of the dissipative stresses \(Q_{\alpha}\), and the boundary dissipation term dissipative stresses on the boundary \(Q_{\alpha}^{(b)}\).

### 3.1 Thermodynamically consistent constitutive relations

Based on previous works [25, 32], the following additive expression of the free energy corresponding to non-local gradient poroplastic materials is adopted

\[
\Psi_s (\varepsilon_{ij}^e, m^e, T, q_{\alpha}, q_{\alpha,i}) = \Psi^e (\varepsilon_{ij}^e, m^e, T) + \Psi_{p,loc} (q_{\alpha}) + \Psi_{p,nloc} (q_{\alpha,i}) \tag{20}
\]

whereby \(\Psi^e\) is the elastic energy of non-isothermal porous media deduced in [22, 4].

\[
\Psi^e = \frac{1}{2} \varepsilon_{ij}^e C_{ijkl}^e \varepsilon_{kl}^e + \frac{1}{2} M (\phi^e)^2 - MB_{ij} \varepsilon_{ij}^e \phi^e - \frac{1}{2} \chi T^2 + \ell \rho_f \phi^e T - A_{ij} \varepsilon_{ij}^e T \tag{21}
\]

Whereas local plasticity contributions to the free energy \(\Psi_{p,loc}\) and non-local contributions to the free energy \(\Psi_{p,nloc}\) are the local and non-local gradient contributions due to dissipative hardening/softening behaviors, which are expressed in terms of the internal variables \(q_{\alpha}\) and their gradient \(q_{\alpha,i}\), respectively.

Once the Coleman’s relationships are deduced from Eq. (17) the following expressions can be obtained

\[
\sigma_{ij} = C_{ijkl}^e \varepsilon_{kl}^e - MB_{ij} \phi^e - A_{ij} T \tag{22}
\]

\[
p = -MB_{ij} \varepsilon_{ij}^e + M \phi^e + \ell \rho_f T \tag{23}
\]

\[
s_s = A_{ij} \varepsilon_{ij}^e - \ell \rho_f \phi^e + \chi T \tag{24}
\]

being \(M\) the Biot’s module [13], \(B_{ij} = b \delta_{ij}\) with Biot coefficient \(b\) the Biot coefficient, and \(C_{ijkl} = C_{ijkl}^s + MB_{ij} B_{kl}\) is the undrained elastic constitutive tensor, whereby \(C_{ijkl}^s\) is the fourth-order elastic tensor which linearly relates stress and strain. Also, \(\chi\) is the porous media heat capacity, \(\ell\) is the latent heat of variation in fluid mass content and \(A_{ij} = \alpha_\theta \delta_{ij}\) is the thermal expansion tensor, with \(\alpha_\theta\) the thermal expansion coefficient.
3.2 Rate form of constitutive equations

Considering the additive decomposition of the free energy potential in Eq. (20) and the flow rule, the following rate expressions are obtained from Eqs. 22-24

\[
\dot{\sigma}_{ij} = C_{ijkl}^{\text{nloc}} \dot{\epsilon}_{kl} - B_{ij} \dot{p} + (B_{ij} \ell \rho f - A_{ij}) \dot{T} - C_{ijkl}^{\text{nloc}} \dot{\lambda} \frac{\partial g}{\partial \sigma_{kl}}
\]

(25)

\[
\dot{p} = -MB_{ij} \dot{\epsilon}_{ij} + M\dot{\phi} + \ell \rho f \dot{T} + MB_{ij} \dot{\lambda} \frac{\partial g}{\partial \sigma_{ij}} - M\dot{\lambda} \frac{\partial g}{\partial p}
\]

(26)

\[
\dot{s}_s = A_{ij} \dot{\epsilon}_{ij} - \ell \rho f \dot{\phi} + \chi \dot{T} - A_{ij} \dot{\lambda} \frac{\partial g}{\partial \sigma_{kl}} + \ell \rho f \dot{\lambda} \frac{\partial g}{\partial p}
\]

(27)

while the evolution law of the local and non-local dissipative stress in Eq. (15) results

\[
\dot{Q}_\alpha = \dot{Q}^{\text{loc}}_\alpha + \dot{Q}^{\text{nloc}}_\alpha
\]

(28)

with

\[
\dot{Q}^{\text{loc}}_\alpha = -\dot{\lambda} H^{\text{loc}}_\alpha \frac{\partial g}{\partial Q_\alpha}
\]

(29)

\[
\dot{Q}^{\text{nloc}}_\alpha = l^2_\alpha \left( H^{\text{nloc}}_{\alpha ij} \dot{\lambda} \frac{\partial g}{\partial Q_\alpha} + \dot{\lambda} H^{\text{nloc}}_{\alpha ij} Q_{ij} \frac{\partial^2 g}{\partial Q_\alpha^2} \right)
\]

(30)

Thereby, local hardening/softening module local hardening/softening module \( H^{\text{loc}}_\alpha \) have been introduced as well as the new non-local hardening/softening tensor non-local hardening/softening module \( H^{\text{nloc}}_{\alpha ij} \) as defined in [25]

\[
H^{\text{loc}}_\alpha = \frac{\partial^2 \Psi^{\text{loc}}_p}{\partial q^2_\alpha}, \quad H^{\text{nloc}}_{\alpha ij} = \frac{1}{l^2_\alpha} \frac{\partial^2 \Psi^{\text{nloc}}_p}{\partial q_{\alpha,i} \partial q_{\alpha,j}}
\]

(31)

\( H^{\text{nloc}}_{\alpha ij} \) is a second order positive defined tensor. The internal characteristic length, \( l_\alpha \), is a physical entity that characterizes the material microstructure [18, 24, 30].

4 A finite element formulation for gradient-based thermo-poro-plasticity

Having established the basic principles of the thermodynamically consistent gradient-based theory for non-isothermal porous media the present section focuses in the formulation of a new \( C_1 \)-continuous FE formulation in order to solve the boundary value problem with the capacity to reproduce both localized and diffuse failure modes that characterized quasi-brittle materials like concretes and soils, see Mroginski and Etse [13], Ripani et al [22] and Coussy [4].
4.1 Incremental formulation

An incremental formulation of the above boundary value problem introduces residual terms, what makes the stress update necessary. The transition from elastic to plastic regimes within a loading step must also be considered. At the end of the \( j + 1 \) iteration of current load step, the incremental equilibrium condition, the fluid mass balance, the heat conservation and the yield condition are studied in a weak form. Thereby, bold symbol for tensors are used instead of the indicial notation employed in previous sections.

\[
\int_{\Omega} \delta \varepsilon^T : \sigma_j d\Omega - \int_{\partial \Omega} \delta u^T t_{j+1} d\partial \Omega = 0 \tag{32}
\]

\[
\int_{\Omega} \delta p \hat{\dot{m}}_{j+1} d\Omega - \int_{\Omega} \nabla \delta p \cdot w_{j+1} d\Omega + \int_{\partial \Omega} \delta p w_{j+1} \cdot n d\partial \Omega = 0 \tag{33}
\]

\[
\int_{\Omega} \delta T \hat{\dot{s}}_{s,j+1} d\Omega - \int_{\Omega} \frac{q_{j+1}}{T_0} \nabla \delta T d\Omega = + \int_{\partial \Omega} \delta T \frac{q_{j+1}}{T_0} \cdot n d\partial \Omega = 0 \tag{34}
\]

\[
\int_{\Omega} \delta \lambda f (\sigma, p, Q_{\alpha})|_{j+1} d\Omega = 0 \tag{35}
\]

In contrast to the local plasticity algorithm, Eq. (35) is not strictly satisfied but in a weak form. Furthermore, it is only fulfilled when the convergence is reached and not necessarily during the iterative process.

Considering the decomposition of the stress tensor in the \( j + 1 \) iteration as \( \sigma_{j+1} = \sigma_j + \Delta \sigma \), and replacing in Eq. (32) it results

\[
\int_{\Omega} \delta \varepsilon^T : \Delta \sigma d\Omega = \int_{\partial \Omega} \delta u^T t_{j+1} d\partial \Omega - \int_{\Omega} \delta \varepsilon^T : \sigma_j d\Omega \tag{36}
\]

Then, the replacement of \( \Delta \sigma \) in the last equation by the linearized form of Eq. (25), results in

\[
\int_{\Omega} \delta \varepsilon^T : (C^s : \Delta \varepsilon - B \Delta p + (B \ell \rho_f - A) \Delta T - C^s : g^s \Delta \lambda) d\Omega = \\
\int_{\partial \Omega} \delta u^T t_{j+1} d\partial \Omega - \int_{\Omega} \delta \varepsilon^T : \sigma_j d\Omega \tag{37}
\]

It can be observed that Eq. (37) is very similar to the incremental equilibrium condition of classical plasticity as it does not include an explicit dependence on the Laplacian of the plastic multiplier.

Considering the incremental decomposition of the infiltration vector \( w_{j+1} = w_j + \Delta w_{j+1} \) and the rate of the fluid mass content \( \dot{m} \), as well as the generalized Darcy’s law for porous media [4], \( w = -k \cdot \nabla p \), the Eq. (33) can be reformulated as
\[
\int_{\Omega} \delta p \left( \frac{\Delta p}{M} + B : \Delta \varepsilon - \frac{\ell \rho_f^T}{M} \Delta T - (B : g^s - g^p) \Delta \lambda \right) \, d\Omega =
- \Delta t \int_{\Omega} \nabla \delta p \cdot k \cdot \nabla p_j \, d\Omega - \Delta t \int_{\Omega} \nabla \delta p \cdot k \cdot \nabla \Delta p \, d\Omega - \Delta t \int_{\partial \Omega} \delta p \, w_{j+1} \cdot n \, d\partial \Omega \quad (38)
\]

Then, considering Fourier law \( q_{j+1} = -\kappa \nabla T \) and the entropy density rate \( \dot{s}_s \) in Eq. (27), the weak form of the heat conservation law in Eq. (34), can be reformulated as

\[
\int_{\Omega} \delta T \left[ (A - \ell \rho_f B) \Delta \varepsilon - \frac{\ell \rho_f^T}{M} \Delta p + \left( \frac{(\ell \rho_f)^2}{M} + \chi \right) \Delta T + (\ell \rho_f B - A) g^s \Delta \lambda \right] \, d\Omega =
- \Delta t \int_{\Omega} \frac{\kappa}{T_0} \nabla \delta T \cdot \nabla \Delta T \, d\Omega - \Delta t \int_{\Omega} \frac{\kappa}{T_0} \nabla \delta T \cdot \nabla T_j \, d\Omega - \Delta t \int_{\partial \Omega} \delta T \cdot \frac{q_{j+1}}{T_0} \cdot n \, d\partial \Omega \quad (39)
\]

Following [18], the yield function \( f \) can be approximated with sufficient accuracy by means of a linear Taylor series around \((\sigma_j, p_j, T_j, Q_\alpha)\),

\[
f (\sigma, p, T, Q_\alpha)|_{j+1} = f (\sigma, p, T, Q_\alpha)|_j + f^s : \Delta \sigma + \ell^p \Delta p + \ell^T \Delta T + f^Q_\alpha \Delta Q_\alpha \quad (40)
\]

Also, from the additive decomposition of the dissipative stress in Eq. (28) it follows

\[
\dot{Q}_\alpha = \dot{Q}^{loc}_\alpha + \dot{Q}^{nloc}_\alpha = -H^l_{\alpha} g^Q_\alpha \dot{\lambda} + l^2_{\alpha} H_n^{loc} g^Q_\alpha \nabla^2 \dot{\lambda} \quad (41)
\]

By replacing Eq. (25) and Eq. (41) into Eq. (40) the weak form of the yield condition is obtained

\[
\int_{\Omega} \delta \lambda \left[ f (\sigma, p, Q_\alpha)|_{j+1} \right] \, d\Omega = \int_{\Omega} \delta \lambda \left[ f (\sigma, p, Q_\alpha)|_j \right] \, d\Omega + \int_{\Omega} \delta \lambda \, f^s : C^s : \Delta \varepsilon \, d\Omega + \int_{\Omega} \delta \lambda \left( \ell^p + \ell \rho_f B - A \right) \Delta T \, d\Omega - \int_{\partial \Omega} \delta \lambda f^s : g^s \Delta \lambda \, d\Omega
+ \int_{\Omega} \delta \lambda f^Q_\alpha \left( -H^l_{\alpha} g^Q_\alpha \Delta \lambda + l^2_{\alpha} H_n^{loc} g^Q_\alpha \nabla^2 \Delta \lambda \right) \, d\Omega = 0 \quad (42)
\]

### 4.2 Galerkin discretization

As it can be observed in Eqs. (37) - (39) and (Eq. (42)) at most first order derivatives of the displacement, pore pressure and temperature fields appear as well as second order derivative of the plastic multiplier. Therefore, displacement, pressure and temperature
fields discretizations require $C_0$-continuous shape functions that are indicated as $N_u$, $N_p$ and $N_T$, respectively. However, $C_1$-continuous shape functions, called $H$, are required for the plastic multiplier discretization. Then, the Finite Element approximations can be expressed as

$$\mathbf{u} = N_u \bar{\mathbf{u}}; \quad p = N_p \bar{p}; \quad T = N_T \bar{T}; \quad \lambda = H \bar{\lambda} \quad (43)$$

where $\bar{\mathbf{u}}$, $\bar{p}$, $\bar{T}$ and $\bar{\lambda}$ are the nodal displacement vector, the pore pressure, the nodal temperature and the plastic multiplier, respectively. Hence considering $\varepsilon = \nabla s \mathbf{u} = \nabla s N_u \bar{\mathbf{u}} = \mathbf{B} \bar{\mathbf{u}}$ and replacing the above entities in Eqs. (37) - (39) and (Eq. (42)) the following set of integral equations is obtained

$$\left\{ \int_{\Omega} \delta \bar{\mathbf{u}}^T \mathbf{B}^T : C^s : \mathbf{B} \, d\Omega \right\} \Delta \bar{\mathbf{u}} - \left\{ \int_{\Omega} \delta \bar{\mathbf{u}}^T \mathbf{B}^T : \mathbf{B} N_p \, d\Omega \right\} \Delta \bar{p}$$

$$+ \left\{ \int_{\Omega} \delta \bar{\mathbf{u}}^T \mathbf{B}^T : (B\ell\rho_j - \mathbf{A}) N_T \, d\Omega \right\} \Delta \bar{T} - \left\{ \int_{\Omega} \delta \bar{\mathbf{u}}^T \mathbf{B}^T : C^s : \mathbf{g}^s \mathbf{H} \, d\Omega \right\} \Delta \bar{\lambda} =$$

$$\int_{\partial \Omega} \delta \bar{\mathbf{u}}^T N_u^{-1} \mathbf{t}_{j+1} \, d\partial \Omega - \int_{\Omega} \delta \bar{\mathbf{u}}^T \mathbf{B} : \sigma_j \, d\Omega \quad (44)$$

$$\left\{ \int_{\Omega} \delta \bar{p}^T \mathbf{B} : \mathbf{B} \, d\Omega \right\} \Delta \bar{\mathbf{u}} + \left\{ \int_{\Omega} \delta \bar{p} \left[ \frac{N_p^T N_p}{M} + \Delta t (\nabla N_p)^T \cdot \mathbf{k} \cdot \nabla N_p \right] \, d\Omega \right\} \Delta \bar{p}$$

$$- \left\{ \int_{\Omega} \delta \bar{p} \frac{\ell\rho_j}{M} N_T \, d\Omega \right\} \Delta \bar{T} + \left\{ \int_{\Omega} \delta \bar{p} N_p \left[ \mathbf{g}^p - \mathbf{B} : \mathbf{g}^s \right] \mathbf{H} \, d\Omega \right\} \Delta \bar{\lambda} =$$

$$- \left\{ \Delta t \int_{\Omega} \delta \bar{p} (\nabla N_p)^T \cdot \mathbf{k} \cdot \nabla N_p \, d\Omega \right\} \bar{p}_j - \Delta t \int_{\partial \Omega} \delta \bar{p} N_p^T w_{j+1} \cdot \mathbf{n} \, d\partial \Omega \quad (45)$$

$$\left\{ \int_{\Omega} \delta \bar{T}^T N_T^T (\mathbf{A} - \ell\rho_j \mathbf{B}) : \mathbf{B} \, d\Omega \right\} \Delta \bar{\mathbf{u}} - \left\{ \int_{\Omega} \delta \bar{T}^T \frac{(\ell\rho_j)^2}{M} N_T \, d\Omega \right\} \Delta \bar{T}$$

$$+ \left\{ \int_{\Omega} \delta \bar{T}^T \frac{\ell\rho_j^2}{M} N_T \, d\Omega \right\} \Delta \bar{\lambda} =$$

$$- \left\{ \Delta t \int_{\Omega} \delta \bar{T}^T \frac{\ell\rho_j}{T_0} (\nabla N_T)^T \cdot \nabla N_T \, d\Omega \right\} \bar{T}_j - \Delta t \int_{\partial \Omega} \delta \bar{T}^T \frac{q_{j+1}}{T_0} \cdot \mathbf{n} \, d\partial \Omega \quad (46)$$
\[ \begin{aligned}
\left\{ \int_\Omega \delta \tilde{H}^T f^s : C^s : \bar{B} \, d\Omega \right\} \Delta \bar{u} &+ \left\{ \int_\Omega \delta \lambda H^T [p^s - f^s : B] N_p \, d\Omega \right\} \Delta \bar{p} \\
+ \left\{ \int_\Omega \delta \lambda H^T [t^T + \ell \rho B - A] N_T \, d\Omega \right\} \Delta \bar{T} &+ \left\{ -\int_\Omega \delta \lambda H^T [f^T : C^s : g^s + \tilde{H}^{loc}_s H_{nloc}^s P] \, d\Omega \right\} \Delta \lambda \\
&= -\int_\Omega \delta \lambda H^T f (\sigma_j, p_j, Q_{\alpha_j}) \, d\Omega 
\end{aligned} \] (47)

where

\[ \nabla^2 (\Delta \lambda) = \nabla^2 (H) \Delta \bar{\lambda} = P \Delta \bar{\lambda} \] (48)

\[ \tilde{H}^s_{loc} = t^Q_{\alpha} H_{nloc}^s g^s_{\alpha} \] (49)

\[ \tilde{H}^n_{loc} = t^Q_{\alpha} H_{nloc}^s g^s_{\alpha} \] (50)

Equations (44)-(47) must hold for any admissible variation of \( \delta \bar{u}, \delta \bar{p}, \delta \bar{T} \) and \( \delta \bar{\lambda} \). Thus, the algebraic equation system in matrix form of the proposed FE formulation for gradient-dependent thermoporoplastic media can be expressed as

\[
\begin{bmatrix}
-K_{ss} & Q_{sp} & Q_{sT} & Q_{s\lambda} \\
Q_{ps} & K_{pp} + \Delta t H_{pp} & Q_{pT} & Q_{p\lambda} \\
Q_{Ts} & Q_{Tp} & K_{TT} + \Delta t H_{TT} & Q_{T\lambda} \\
Q_{s\lambda} & Q_{p\lambda} & Q_{\lambda T} & -K_{\lambda \lambda}
\end{bmatrix}
\begin{bmatrix}
\Delta \bar{u} \\
\Delta \bar{p} \\
\Delta \bar{T} \\
\Delta \lambda
\end{bmatrix}
= \begin{bmatrix}
F_{s}^{\text{int}} - F_{s}^{\text{ext}} \\
-F_{p} \\
-F_{T} \\
-F_{\lambda}
\end{bmatrix} \] (51)

Submatrices of Eq. (51) were obtained by inspection from Eqs. (44)-(47). Also, in Table 1 the solution algorithm of the boundary value problem is summarized.

The main and most important difference between this selective \( C_1 \)-continuous FE formulation and the one based on \( C_0 \) continuity approximations for gradient plasticity proposed by [26, 31] is the solution procedure. While present formulation requires only the solution of Eq. (51), the FE approaches proposed by the aforementioned authors require an additional global iteration to obtain the plastic multiplier.

REFERENCES


1) Compute matrices of Eq. (51)
2) Solve the algebraic system of Eq. (51) in terms of the increments $\Delta \bar{u}$, $\Delta \bar{p}$, $\Delta \bar{T}$ and $\Delta \bar{\lambda}$
3) Update primary variables $\Delta \bar{u}_{j+1} = \Delta \bar{u}_j + \Delta \bar{u}$,
   $\Delta \bar{p}_{j+1} = \Delta \bar{p}_j + \Delta \bar{p}$, $\Delta \bar{T}_{j+1} = \Delta \bar{T}_j + \Delta \bar{T}$ and $\Delta \bar{\lambda}_{j+1} = \Delta \bar{\lambda}_j + \Delta \bar{\lambda}$
4) On each integration point compute:
   $\Delta \varepsilon_{j+1} = B \Delta \bar{u}_{j+1}$
   $\Delta \lambda_{j+1} = H \Delta \bar{\lambda}_{j+1}$
   $\nabla^2 (\Delta \lambda_{j+1}) = P \Delta \bar{\lambda}_{j+1}$
   $q_{\alpha_{j+1}} = q_{\alpha_0} + g^Q_{\alpha} \Delta \lambda_{j+1}$
   $\nabla^2 q_{\alpha_{j+1}} = \nabla^2 q_{\alpha_0} + g^Q_{\alpha} \nabla^2 (\Delta \lambda_{j+1})$
   $\sigma^t = \sigma_0 + C^s : \Delta \varepsilon_{j+1} - B N_p \Delta \bar{p}_{j+1}$
   IF $f (\sigma^t, q_{\alpha}, \nabla^2 q_{\alpha})_{j+1} > 0$
   $\sigma_{j+1} = \sigma^t - \Delta \lambda_{j+1} C^s : g^s$
   ELSE
   $\sigma_{j+1} = \sigma^t$
   END
5) Check convergence criterion, i.e. balance between internal and external energy. If it is not achieved go to 1

Table 1: Gradient-thermo-poro-plasticity algorithm for $C_1$-continuous FE


MODELLING FLUIDISATION AND SEDIMENTATION USING MATERIAL POINT METHOD

M. MARTINELLI\textsuperscript{1} and A. ROHE\textsuperscript{2}

\textsuperscript{1}\textit{Deltares}  
Delft, The Netherlands  
mario.martinelli@deltares.nl

\textsuperscript{2}\textit{Deltares}  
Delft, The Netherlands  
alex.rohe@deltares.nl

\textbf{Key words:} Material Point Method, Breaching, Sedimentation, Fluidisation

\textbf{Abstract.} Fluidisation and sedimentation processes occur in the contact zone between soil and free water where a transition between free water and soil skeleton occurs. In order to numerically model these phenomena the material point method (MPM) is used. In this study, a formulation of the material point method is exploited, which uses two sets of material points for soil and water respectively. In such way, not only the soil-water interaction can be described, such as the accumulation and dissipation of excess pore pressures, but also the transition from free water to groundwater as well as fluid-like behaviour of soil typical for fluidisation and sedimentation problems can be simulated. In this paper, examples of a collapsing submerged sand column are presented.

\section{INTRODUCTION}

Modelling of fluidisation and sedimentation processes is an essential task in many areas of environmental and civil engineering. The coastal safety which involves the stability of submerged slopes is an important issue in many deltaic areas. In particular, in the Netherlands the problem has a great impact in the province of Zeeland, in the south-east of the country, characterized by numerous islands, where the shoreline has been severely damaged by sea attack, and submarine landslides compromise the safety of the area. This phenomenon needs to be thoroughly investigated in order to enforce the design of mitigation measures.

Fluidisation and sedimentation occurs in the contact zone between soil and free water, where the conventional concepts of soil mechanics, related to the existence of a soil skeleton, do not longer apply and there is a transition zone between free water and solid
skeleton. Such processes may occur in case of failure of submerged slopes, and the prediction of the new configuration of the slope is strongly related to the ability to correctly simulate both the failure mechanism together with fluidisation-sedimentation processes.

One approach for investigating soil-water interaction is to perform physical experiments, but usually they are costly even though at small scale. The advent of computers and computational mechanics provided a more convenient way to simulate such problems. The models developed within the framework of computational fluid dynamics (CFD), such as *Delft3D* [6], are mainly focused on sediment flow and deposition, but they are less suitable to study the mechanical response of soil during and after the sedimentation process. In order to capture these phenomena, a numerical method which can describe both water and soil with high level of accuracy is needed.

During the last two decades, a large number of point-based computational methods were developed to simulate large deformation problems together with solid-liquid interaction. These methods are known as meshless or meshfree methods, as a set of points is used to calculate the approximate solution independent, or in a combination with a background mesh, which is regularly used in the classical approach of finite element method (FEM) or finite difference method (FDM). For instance, smoothed particle hydrodynamics (SPH) [12, 13, 5, 17], particle finite element method (PFEM) [16, 15, 14] and material point method (MPM) [9, 18] are some examples which can be used to approximate both solid and liquid behaviour together with their mutual interaction.

The aim of this paper is to describe recent advances of MPM for fluid-solid interaction problems. In the following section the general principles of MPM with focus on liquids and granular materials are outlined. Next, the basic continuum equations and the discretised formulation are briefly described. Lastly, an example of soil-water interaction problem modelled with MPM involving fluidisation and sedimentation phenomena is illustrated.

2 GENERAL CONCEPTS

2.1 History of material point method

The original formulation of MPM was developed by Harlow (1964) for fluid mechanics and later applied to solid mechanics [18] and dry granular materials [20, 21, 22]. Later, the method was extended to handle saturated soils [10] with a numerical approach following the so called v-w formulation [24]. This formulation, which uses the velocity of both solid and liquid constituent as the primary unknowns, was applied to several small and large deformation problems and is able to capture the physical response of saturated soil under dynamic loading [7]. However, only one set of material points was used for both the solid and the liquid phase; therefore groundwater flow and the transition between free water and groundwater could not be captured as well as fluid-like behaviour of the soil, typical of fluidisation and sedimentation problems. Recently, a formulation with two sets of material points (so called *two-layer formulation*) was proposed [2, 1, 19] to overcome such difficulties.
2.2 The basics of standard MPM

MPM is an advancement of the finite element method particularly suited for large deformations. In MPM the continuum body is represented by a set of Lagrangian points, called material points (MPs). Large deformations are modelled by MPs moving through a background mesh, which also covers the domain where the material is expected to move. The MPs carry all physical properties of the continuum such as stresses, strains, density, momentum, material parameters and other state parameters, whereas the background mesh is used to solve the balance equations without storing any permanent information.

In this study, an implementation using an explicit integration scheme [10] is applied. For each time step, the momentum balance equation is solved at the nodes of the background mesh resulting in the velocity field. After computing the strain increment at the location of the material points, the updated stress is calculated and the mesh is reset to its original configuration. The complete procedure for a single time step is illustrated in Figure 1.

![Figure 1](image1.png)

**Figure 1:** (a) Configuration of material points and background mesh at the beginning of the calculation. (b) Deformed configuration after solving the balance equation. (c) Reset the background mesh to the initial configuration and update the position of MPs

2.3 The basics of two-layer formulation

Figure 2 shows an example of a submerged slope which is characterized by two materials: saturated soil and free water. The soil is composed of two constituents, solid grains and groundwater respectively, whereas the free water is pure liquid. In general, the water can flow from the soil body into the water reservoir or vice-versa, and interacts with the solid skeleton through drag forces. Moreover, the soil can behave as a solid porous media with liquid in pores (solid-like response) or as in liquefied state in which soil grains float in the liquid (liquid-like response). In such conditions, the use of a single material point to simulate the behaviour of both the soil and free water appears not possible and thus two sets of material points were introduced [2, 1, 19]: one describes the physical behaviour of the liquid constituent and the other one the solid.

The motion of both sets of material points is described by the system of momentum balance equations, using separate velocity fields, \( \mathbf{v}_S \) and \( \mathbf{v}_L \), for solid and liquid constituent respectively:
\[ \nabla \cdot \sigma' S + n_S \nabla \cdot \sigma_L + \bar{\rho}_S g + f_d = \bar{\rho}_S v_S \]  
\[ n_L \nabla \cdot \sigma_L + \bar{\rho}_L g - f_d = \bar{\rho}_L v_L \]

The symbols \( \bar{\rho}_S \) and \( \bar{\rho}_L \) represent the partial densities of the solid and liquid constituent, defined as the ratio of the mass with respect to the reference volume; \( n_S \) and \( n_L \) are the concentration ratio's for the solid and liquid constituent, defined as the ratio of the partial density with respect to the corresponding mass density. Note that \( n_L \) is equal to 1 in case of pure liquid and \( n_S = 1 - n_L \) in case of fully saturated soil. The stress tensors are stated as \( \sigma'_S \) for solid (effective stress) and \( \sigma_L \) for liquid. The vector of gravity is \( g \), and \( f_d \) represents the drag force vector exerted by the liquid on the solid part. The drag force is calculated with the Darcy equation as

\[ f_d = n_L \frac{\mu}{\kappa} (v_L - v_S) \]  

where \( \mu \) and \( \kappa \) are the liquid dynamic viscosity and the soil intrinsic permeability respectively. The Kozeny-Carman formula [3] is used to update the soil intrinsic permeability as follows:

\[ \kappa = \frac{D_p^2}{A} \frac{n^3}{(1 - n)^2} \]  

where \( D_p \) is the grain size diameter, \( n \) is the soil porosity and the constant \( A \) is set to 150 according to Ergun [8].

An important concept of this formulation is represented by the State, of the soil-liquid mixture which can be Solid or Liquid, as shown in Figure 3. On the left side, the mixture is characterized by a low porosity: the grains of the solid skeleton are in contact and the behaviour can be described by constitutive models designed for granular materials. In this condition the State is defined as Solid. Conversely, on the right side, the mixture has a high porosity, the grains are not in contact and float together with the liquid phase. In this condition the effective stresses are equal to zero (\( \sigma'_S = 0 \)) and the response of the
mixture is described by the Navier-Stokes equation. In this condition the State is defined as Liquid.

In order to distinguish between the two aforementioned states, an additional parameter is used in the formulation, i.e. the maximum porosity $n_{\text{max}}$. During the fluidisation process the mixture with initially Solid State is characterized by two aspects: the mean effective stress decreases and the porosity increases. At a particular moment, the mean effective stress may become nil when the contact forces between the grains vanish. However, the fluidisation occurs only if the grains are significantly separated, so that the porosity is larger than $n_{\text{max}}$ and the State turns to Liquid. In the reverse process, i.e. the sedimentation of a fluidized mixture, the porosity decreases due to the fact that the solid grains get closer to each other. However, the effective stresses recur only if the porosity is smaller than $n_{\text{max}}$, i.e. the grains are close enough to be in contact, and the State turns to Solid.

Figure 3: (a) Solid-liquid mixture with low porosity (Solid State); (b) Solid grains floating in the liquid in case of high porosity (Liquid State)

In case of Solid State, the effective stress rate in solid MPs is calculated using the constitutive law. In case of Liquid Status, the rate of effective stress in solid MPs is set to zero.

The stress rate for the liquid phase is computed using:

\[ \dot{\sigma}_{\text{dev},L} = 2\mu_{\text{eq}}\dot{\varepsilon}_{\text{dev},L} \quad \text{with } p_L > 0 \]  

where $\dot{p}_L$ and $\dot{\sigma}_{\text{dev},L}$ are the mean and deviatoric part of the stress rate tensor of the liquid constituent, $K_L$ is the bulk modulus of the liquid constituent, $\mu_{\text{eq}}$ is the equivalent viscosity which is assumed to be dependent on the volume of solid grains suspended in the water [4], and $\dot{\varepsilon}_{\text{dev},L}$ is the deviatoric part of the strain rate tensor.

The volumetric strain rate $\dot{\varepsilon}_{\text{vol},L}$ of the liquid constituent is calculated according to the mass balance equation as follows:

\[ \dot{\varepsilon}_{\text{vol},L} = n_L(\nabla \cdot \mathbf{v}_{L}) + n_S(\nabla \cdot \mathbf{v}_S) \] (7)
Equation (6) is used only in case of **Liquid State**, i.e. pure liquid or fluidized mixture. In case of **Solid State**, the shear stresses in the liquid part vanish.

### 3 DISCRETISED FORM

First, the conservation equations (1) and (2) are transformed into the weak form. The momentum equations are multiplied by the corresponding virtual velocity \((\delta v_S, \delta v_L)\) and integrated over the current domain \(\Omega\). The virtual velocities \((\delta v_S, \delta v_L)\) have to be continuous over the domain \(\Omega\) and nil on the prescribed displacement boundary \((\delta \hat{v} = 0\) on \(\partial \Gamma_u\)).

Using the Gauss' theorem with the Voigt notation, the variational formulation of Equations (1) and (2) can be written as follows:

\[
\begin{align*}
\int_{\Gamma} \delta v_S t_S \cdot n \ d\Gamma - \int_{\Omega} \nabla (\delta v_S^T \sigma_S) \ d\Omega + \int_{\Gamma} \delta v_S n_S t_L \cdot n \ d\Gamma \\
- \int_{\Omega} \nabla (\delta v_S^T n_S) \sigma_L \ d\Omega + \int_{\Omega} \delta v_S^T \rho_S g \ d\Omega + \int_{\Omega} \delta v_S^T n_L^2 \mu/\kappa (v_L - v_S) \ d\Omega \\
- \int_{\Omega} \delta v_S^T \rho_S \dot{v}_S \ d\Omega = 0
\end{align*}
\]

Using the same shape function matrix for the virtual velocities \(\delta \hat{v} = N \delta \hat{v}\), the discretized form yields:

\[
\begin{align*}
\mathbf{F}_{ext,t}^S + \mathbf{F}_{ext,g}^S - \mathbf{F}_{int}^S - \mathbf{F}_{n}^S + \mathbf{F}^d = \mathbf{M}_S \dot{\hat{v}}_S & \quad (11) \\
\mathbf{F}_{ext,t}^L + \mathbf{F}_{ext,g}^L - \mathbf{F}_{int}^L - \mathbf{F}_{n}^L - \mathbf{F}^d = \mathbf{M}_L \dot{\hat{v}}_L & \quad (12)
\end{align*}
\]

with
\[ F_{ext,t}^S = \int_{\Gamma} N^T (t_S + n_S t_L) \cdot n \, d\Gamma \; ; \quad F_{ext,t}^g = \int_{\Omega} \bar{\rho}_S N^T g \, d\Omega \]
\[ F_{int}^S = \int_{\Omega} B^T (\sigma' + n_S \sigma_L) \, d\Omega \; ; \quad F_{ext,t}^g = \int_{\Omega} N^T \nabla n_S \sigma_L \, d\Omega \]
\[ F^d = (\hat{v}_L - \hat{v}_S) \int_{\Omega} n_L^2 \mu / \kappa N^T N \, d\Omega \; ; \quad M_S = \int_{\Omega} \bar{\rho}_S N^T N \, d\Omega \]
\[ F_{ext,t}^L = \int_{\Gamma} N^T n_L t_L \cdot n \, d\Gamma \; ; \quad F_{ext,t}^g = \int_{\Omega} \bar{\rho}_L N^T g \, d\Omega \]
\[ F_{int}^L = \int_{\Omega} B^T n_L \sigma_L \, d\Omega \; ; \quad F_{ext,t}^g = \int_{\Omega} N^T \nabla n_L \sigma_L \, d\Omega \]
\[ M_L = \int_{\Omega} \bar{\rho}_L N^T N \, d\Omega \]

where \( F_{ext,t}^S \) and \( F_{ext,t}^g \) are the vector of nodal external forces, \( F_{int}^S \) is the vector of nodal internal forces, \( F^d \) is the vector of solid-liquid interaction forces related to the gradient of the concentration ratio, \( F_{ext,t}^d \) is the vector of nodal drag forces related to the Darcy law and \( M \) is the nodal mass matrix. Indices \( S \) and \( L \) refer to the constituent.

4 SUBMERGED COLUMN OF SAND

An example which highlights the main aspects of the fluidisation and sedimentation process is the collapse of a submerged column of sand. The geometry is shown in Figure 4a. A standard linear elastic perfectly plastic constitutive model with a Mohr-Coulomb failure criterion is used for the solid skeleton and a standard Newtonian compressible constitutive model is used for water. The parameters for both constitutive models are summarized in Table 1. Since Equations (11) and (12) are solved with an explicit time integration scheme, the bulk modulus of the water is reduced by a factor of 100 in order to reduce the computational time.

A quasi plane strain condition is assumed for the simulation, considering only a strip of tridimensional elements. The mesh is shown in Figure 4b and consists of 5760 4-node tetrahedral elements. At the beginning of the calculation, 8 material points per element (4 liquid and 4 solid) are assigned to the soil domain whereas 4 liquid material points are used for the free water. At all boundaries, the displacements are constrained in perpendicular direction and free in the longitudinal one.

The soil is assumed to be fully saturated. The initial stress in the water is set as hydrostatic and the effective stress in the solid skeleton is set as:

\[ \sigma'_{S,zz}(t=0) = (\bar{\rho}_{S,0} - \rho_L) g (z - z_{top}) \]
\[ \sigma'_{S,xx}(t=0) = \sigma'_{S,yy}(t=0) = K_0 \sigma'_{S,zz}(t=0) \]

where \( \bar{\rho}_{S,0} \) is the mass concentration for the soil, \( z_{top} \) is the value of \( z \) at the top of the soil column and \( K_0 \) is a constant equal to 0.5.
The results are shown in Figure 5 for several phases during collapse. Only information at the position of the solid MPs is represented in order to generate a figure readable. The porosity at the location of the material points is plotted for $t = 0.2, 1, 3$ and 6 seconds, and can vary approximately from 0.43 to 1.0. However, the legend has a smaller interval to focus on the range of solid-like behaviour ($n < 0.5$).

At time $t = 0 \, s$, the initial porosity is homogeneous with the value of 0.43 and the initial mean effective stress is not nil in any material point. As the time proceeds, the effective stress tends to decrease and drops to zero at the top and at the right boundary of the soil column, where the soil is in contact with the free water.

Due to the hydro-mechanical coupling, the positive value of the dilatancy angle induces negative excess pore water pressures $\Delta p_L$ (extension) that initially prevent any volume increase in the solid skeleton and increases the mean effective stress. However, these $\Delta p_L$ are rapidly dissipated in the vicinity of the transition zone, such that the porosity of the solid MPs increases.
As the calculation continues, the porosity of the outer layer exceeds the maximum porosity and consequently the soil fluidises. The thickness of this fluidised zone progressively increases and solid MPs, which are part of it, move downwards towards the base of the model. This phenomenon is known as breaching. It is worth noticing that the column is standing up for almost 3 s before the global failure starts.

![Porosity of solid MPs for collapsing submerged sand column.](image)

The effect of the soil permeability is investigated by using a larger value of $D_p$ for which the results are shown in Figure 6a. As the soil is more permeable, the dissipation of excess pore pressures $\Delta p_L$ generated as a result of the dilatancy in saturated soils is much faster and the effective stress in the solid MPs drops rapidly to zero, so that the stand-up time is much shorter and the column collapses after a few seconds.

The effect of the dilatancy angle is also studied by setting the value of $\psi$ to zero for which the results are shown in Figure 6b. In this case, the porosity remains almost constant with the exception of a thin layer which is in contact with the free water. Moreover, the breaching phenomenon seems not to play a significant role as it is in the reference analysis, but rather a dominating plastic deformation is observed.

5 CONCLUSION

The paper shows the potential of the material point method to model several aspects of solid-fluid interaction problems involving large displacements and including complex
phenomena such as fluidisation and sedimentation. The solid-liquid mixture is treated as saturated soil or fluidized granular material, and the liquid flow is described either by the Darcy law or the Navier-Stokes equation. The problem of submerged soil column is successfully modelled and, although a simple constitutive model for soil is used, the complex phenomena such as breaching seems to be captured. The influence of the dilatancy angle and the soil permeability is also partly investigated.

It can be concluded that MPM is a promising tool in modeling solid-fluid interaction problems and, in the near future, can be also used to improve construction processes in the dredging industry.

6 ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. PIAP-GA-2012-324522 MPM-DREDGE. The authors would like to thank Prof. Pieter A. Vermeer and Dr. Lars Beuth for their contributions to the research efforts leading to this publication.

REFERENCES


RECENT DEVELOPMENT IN THE COMBINED FINITE-DISCRETE ELEMENT METHOD

ESTEBAN ROUGIER¹, EARL E. KNIGHT¹, ZHOU LEI¹ and ANTONIO MUNJIZA²

¹ Geophysics Group (EES-17), Los Alamos National Laboratory
P.O. Box 1663, 87545, New Mexico, United States
knighte@lanl.gov
zlei@lanl.gov
erougier@lanl.gov
http://www.lanl.gov

² Department of Engineering, Queen Mary, University of London
Mile End Road, London, E1 4NS, United Kingdom
a.munjiza@qmul.ac.uk
http://www.qmul.ac.uk

Key words: DEM, FEM, Combined Finite-Discrete Method (FDEM), Contact, Fracture.

Abstract. Discontinua modeling techniques, such as Discrete Element Method (DEM), Discontinuum Deformation Analysis (DDA), Combined Finite-Discrete Element Method (FDEM) and Numerical Manifold Method (NMM) have become important analysis tools within the Computational Mechanics field. These methods can now be grouped as methods of Computational Mechanics of Discontinua.

The Combined Finite-Discrete Element Method bridges the gap between Finite and Discrete Element Methods. As such, it has become a tool of choice for problems involving fracturing, fragmenting and complex shapes. The key advantage of FDEM is the introduction of finite displacements, finite rotations, and finite strain based deformability combined with suitable material laws; these are then merged with discrete element-based transient dynamics, contact detection, and contact interaction solutions and objective discrete crack initiation and crack propagation solutions that have a great deal of fidelity in reproducing complex fracture patterns and eventual fragmentation.

After nearly 25 years of development, the method has been successfully applied in rock mechanics, biomedical engineering, structure engineering and mechanical engineering. This paper summarizes the most recent development efforts in FDEM.

1 HOSS

Engineers have encountered discontinua in their practice for a long time. A classic example is the existence of discontinuities in rock masses. Shi was one of the pioneers in attempting to address the simulation of these features via the Discontinuous Deformation Analysis (DDA), Shi¹ Shi & Goodman². DDA was at first developed for static problems and later on extended to dynamic problems. In parallel, Cundall³ developed the Distinct Element Method (DEM) for dynamic problems, which was extended by many other researchers (Cundall⁴,
Cundall & Strack\cite{5}, Hocking et al.\cite{6,7} and Mustoe et al.\cite{8,9}). For problems involving transitions between continuum and discontinuum, i.e., fracture and fragmentation, Munjiza\cite{10} introduced the Combined Finite Discrete Element Method (FDEM). The FDEM was applied to many different problems (Munjiza et al.\cite{11,12}, Munjiza\cite{13}, Munjiza & Andrews\cite{14}, Munjiza & John\cite{15}, and Munjiza & Latham\cite{16}). All these methods, which can be grouped under the Computational Mechanics of Discontinua umbrella, have now become mainstream computational and simulation tools with applications ranging from original rock mechanics to nanotechnology, material sciences, oil exploration, mineral processing, medical engineering and many other fundamental and applied disciplines. The first textbook on the topic was published by Wiley in 2011 (Munjiza et al.\cite{17}) and it presents a comprehensive attempt to compile the latest algorithmic advances made in the field, with a primary focus on those being used in FDEM.

FDEM bridges the gap between Finite and Discrete Element Methods. The key advantage of FDEM is the introduction of finite displacements, finite rotations, and finite strain based deformability combined with suitable material laws; these are then merged with discrete element based transient dynamics, contact detection, and contact interaction solutions and objective discrete crack initiation and crack propagation solutions that have a great deal of fidelity in reproducing complex fracture patterns and eventual fragmentation.

At Los Alamos National Laboratory, research and development of a FDEM software package called the Hybrid Optimization Software Suite (HOSS) is well under way. HOSS (Knight et al.\cite{19} and Rougier et al.\cite{20}) is an original discontinua simulation platform whose main purpose is to bring computational science of discontinua closer to applied scientists or practicing engineers, who may not be experts in details of the complex simulation procedures involved. HOSS includes a large strain large displacement finite element formulation that is addressed in depth in Munjiza et al.\cite{18}. This is complemented by a patented novel fluid solver which is integrated into the code.

HOSS is aimed at internationalizing research and development work in many applications of science of discontinua, such as: Material Science, Laboratory Testing, Occupant Risk Analysis, Mining, Seismic Analysis, Blasts & Impacts, Oil and Gas Proppants and Fracking, Medical Engineering, and many others. The goal is to provide the academic community and general users with a parallel 2D/3D FDEM capability that ultimately leads to commercialization of a more robust platform for International distribution and use. In the rest of the paper, the most recent development efforts for FDEM that are incorporated in HOSS are summarized.

2 REAL SHAPE PARTICLES

The performance evaluation of the contact interaction and contact detection algorithms has been achieved with the help of model shown in Figure 1, which consisted of a cubical, hollow raster of particles of general shapes placed inside a rigid spherical container (Rougier et al.\cite{21}). In this case, four different types of particles of general shapes were selected and the raster is centered in a rigid spherical container. There is no initial overlap between the particles. Each particle is given an initial velocity of 100 m/s, pointing towards the center of the container. The system is allowed to relax, and as a result, a random pattern of collisions takes place between the particles, as shown in Figure 1.
3 GRAND SCALE PARALLELIZATION

Recent work on HOSS has demonstrated that it is possible to build a problem-specific (as opposed to computer architecture-specific) virtual parallel machine for parallelization of existing FDEM software packages (Lei et al. [29]). The developed virtual parallel machine consists of four virtual engines, one of which is the unchanged existing sequential FDEM package. Porting the parallel software to different parallel computer architectures is thereby reduced to implementing some optimized aspects of the three remaining virtual engines. The performance, especially efficiency of the parallelized software still depends on the particular architecture however, the test results shown for both 2D (Figure 2) and 3D (Figure 3) clearly demonstrate that it is possible to achieve very high efficiency even when a relatively large number of processors are employed.

Figure 1: Time evolution of collapsing raster.

Figure 2: HOSS 2D parallelization. Left: Particle mechanics raster. Center: Efficiency as a function of the # of processors. Right: Speed-up as a function of the # of processors.

Figure 3: HOSS 3D parallelization. Top Left: 3D raster of 126×126×126 elastic tetrahedron. Top Right: Efficiency as a function of the # of processors.. Right: Speed-up as a function of the # of processors.
4 NEXT GENERATION OF VALIDATION AND VERIFICATION

Shale Under Shear Experiment. Developing a class of simple benchmark problems for discontinua simulations has become a must. Problems with simple analytical solutions should be used. In addition, experimental problems with a relatively standard set up should be employed. For instance, in Figure 4 a FDEM model was used to simulate the fracture process of shale under a direct shear experiment. The shale sample is compressed in an asymmetric fashion by the action of two semi-circular, rigid anvils. The focus was to characterize the behavior of the generic shale sample and compare this with the experimental observations (Carey et. al.[22]). FDEM numerical simulations were able to replicate the main features of the fracturing processes while showing the importance of fluid penetration into fractures as well as layering in determining the final fracture patterns.

\[ \text{Figure 4:} \text{ 2D Simulation of Laboratory Test: Left: Experimental results. Right: Simulation results.} \]

Split Hopkinson Pressure Bar Experiment. Another classic example of FDEM laboratory replication is shown in Figure 5 where a 3D Split Hopkinson Pressure Bar (SHPB) simulation is displayed (Rougier et al.[23]).

\[ \text{Figure 5:} \text{ 3D Simulation of SHPB Experiment: Left: Experimental results. Center: Simulation results. Right: Equivalent tensile stress as a function of time.} \]
5 DIVERSE FIELD OF APPLICATIONS

Acoustic Emissions Related to Damage. Validating different aspects of discontinua simulations does require custom designed experiments, thus necessitating close collaboration between experimental and simulation approaches. For example, the modeling of nonlinear effects caused by damage as seen in Figure 6 appears to be an enticing area for the usage of FDEM.

Figure 6: Acoustic Emission Simulation Example. HOSS simulation results shown for S₀.

Glass Fracture. A shell element formulation has been also included into the HOSS package. This type of element is useful for the simulation of glass fracturing, as shown in Figure 7, where a coin-shape projectile impacts a rectangular glass panel that is supported at the ends by a set of metal stripes (green stripes). The fractures generated inside the glass panel are for an impact velocity of 3.0 m/s (Munjiza et al. [24]). As an example, this unique capability can benefit occupant risk analysis.

Figure 7: An example of glass fracture – 2.5D shell problem.

Fragmentation. The system shown in Figure 8 has been used to demonstrate the application of HOSS to the simulation of rock blasting. A 3D bench with a four explosive boreholes was modeled. The model itself is a 3D extension of the 2D Ning et al. [25] bench model. The 3D mesh was approximately 3.5M elements. The explosives are detonated and as
the rock material breaks a myriad of fragments are created.

**Figure 8**: 3D FDEM Bench model. Left: Longitudinal cut of a borehole, Right: Full view 4 exploding boreholes.

**Block Caving.** Block caving is another example of a mining technique that can be simulated using HOSS. The model and results for this case are shown in Figure 9. More details on HOSS applicability to mining can be found in Rougier et al.\[26\].

**Figure 9**: 3D FDEM Drawbell model. Left: Geometry. Center: Initial sequence, longitudinal cut of HOSS 3D results. Right: Latter sequence, longitudinal cut of HOSS results.

**Seismic Loads.** Another field of application for the HOSS package is the simulation of the response of masonry structures to seismic loads. For demonstration purposes, an example of the response of a dome to a horizontal (shaking) seismic load is shown in Figure 10.

**Figure 10**: 3D Setup of the FDEM Drawbell model.
The approach to this diverse application of FDEM (historical monument preservation) can be found in De Lorenzi-Venneri et al.[27]. Over the next few years FDEM researchers will work with University of California, San Diego students to analyze the effects of structural strengthening to the Duomo of the Cathedral of Santa Maria del Fiore, located in Florence, Italy.

**Impact.** Another intriguing application wherein FDEM can be being applied is in the arena of seismic efficiencies. With efforts to emplace seismometers on Mars in the near future the drive is on to find a simulation tool that can replicate bolide impacts. Accurate free field impact wave forms must be transferred to seismic codes and to date researchers are finding severe limitations in several continuum and Eulerian approaches. As an example HOSS was used to simulate a high speed bolide impacts, results are shown in Figure 11.

![Figure 11: Example of an oblique impact of a spherical bolide into a disk of unconsolidated material. The view has been cut to show the wave propagation through the target material.](image)

**Fracking.** The accurate simulation of hydrofracking continues to be a challenge for numerous codes for a variety of reasons. To enhance efficiencies and accuracy, LANL researchers have incorporated an integrated solid-fluid solver (ISF) within the framework of HOSS, Knight et al.[28]. The ISF will account for fluid flow through fracturing porous solid in 2D/3D, fluid flow through crack manifolds, pressure wave propagation through fluid, fluid-solid interaction, and dynamic hydrofracture. It fully integrates the solid & fluid aspects of the complete fracture problem by handling failure/fracturing in a continuum-discontinuum regime.

![Figure 12: Illustration of HOSS borehole fracking. Left: 400m depth. Right: 1000m depth.](image)
The fluid phase is captured within the same grid of the solid phase using a modified Eulerian formulation. This therefore eliminates the need of continuously mapping physical variables between the fluid and solid domains. In essence, the ISF resolves the problem explicitly with a fluid phase time step which is similar in nature to the solid phase time step. For HOSS, the ISF is now parallel for 2D simulation space with 3D coming in the next few months. The relevance of accurate in-situ borehole hydrofrac in a shale material can be seen in the two HOSS simulations shown in Figure 12.

**Proppants.** Industry experts estimate that for a single $10 million dollar lateral wellbore fracking operation, only 10% of the hydrocarbons contained in the rock are extracted. Many simulation tools are being developed to address hydrofrac processes, but very few can accurately address the near wellbore interactions, especially the role proppants play. A research pathway now exists wherein the characterization and upscaling of the behavior of millimeter-size proppant grains can be accomplished. The goal would be to utilize a mesoscale thermo-hydro-mechanical (multi-physics) continuum discontinuum (FDEM) fracture-based simulation tool to understand the role of proppant in improving recovery efficiencies and lowering costs. Essentially, micro-scale proppant modeling results will be used to generate constitutive laws for HOSS (Figure 13).

![Illustration of HOSS proppant upscaling.](image)

**Blood Flow.** There is a diverse field of problems where discontinuous solids are immersed in a fluid media. A classic example is the case of red blood cells immersed in blood plasma. A conventional approach to model such a fluid is to use continuum-based mechanics techniques in conjunction with non-Newtonian fluid formulation. The problem with this approach is the continuum assumption; for instance when applied to modeling a mechanical heart, it fails to predict the damage caused to red blood cells, which simply become smeared in the continuum formulation. An alternative discontinuum formulation to the same problem is shown in Figure 14. Not only does it take the discrete nature of red blood cells into account, but also it represents the deformation and other phenomena associated with red blood cells. Nevertheless the problem with discontinua formulations are its algorithmic complexity, validation and verification complexities, and the grand scale nature of such simulations (computational resources). Nevertheless, LANL researchers have identified a pathway wherein HOSS blood flow simulations can be upscaled to create perturbed continuum laws. Because of its industrial significance, this type of problems have the immense potential to led to advancements in toroidal flow analysis near heart pumps.
6 CONCLUSIONS

- The HOSS integrated FDEM simulation platform comprising state of the art recent developments has been introduced and demonstrated using various applications.
- HOSSedu is in beta testing and is expected to be released in the coming months. It is a fully parallel 2D/3D FDEM dry fracture platform.
- HOSS is aimed at attempting to bridge the gap between applications and simulations. Ultimately, any numerical tool that can facilitate multidisciplinary and interdisciplinary work at the theoretical, experimental and industrial scales will enhance collaborations worldwide.

REFERENCES


STRUCTURAL BEHAVIOR OF UNREINFORCED MASONRY PILES OF DIFFERENT THICKNESSES SUBJECT TO COMPRESSION LOADS

NELSON D. AGÜERA¹, MIGUEL E. TORNELLO² AND CARLOS D. FRAU³

¹CEREDETEC (Centro Regional de Desarrollos Tecnológicos para la Construcción, Sismología y la Ingeniería Sísmica). Universidad Tecnológica Nacional Facultad Regional Mendoza. Rodríguez 273, Ciudad, Mendoza, Argentina (5500) naguera@frm.utn.edu.ar
²³CEREDETEC (Centro Regional de Desarrollos Tecnológicos para la Construcción, Sismología y la Ingeniería Sísmica). Universidad Tecnológica Nacional Facultad Regional Mendoza. Rodríguez 273, Ciudad, Mendoza, Argentina (5500) mtornell@frm.utn.edu.ar, cdfrau@frm.utn.edu.ar

Key words: Masonry, Structure, FEM.

Abstract. Many historic buildings emplaced in seismic regions have as structure main, unreinforced masonry walls with significant thicknesses. Research works oriented to evaluate the structural response of these constructions discard the thickness of the walls in the structural response of the building. This paper studies the influence of the thickness of the wall under compressive loads on piles of unreinforced masonry. The walls tested with this thickness correspond to those actually found in historic buildings located on the province of Mendoza, Argentina. These buildings are currently in service and do not have an earthquake resistant structure that meets current design regulations. The tested piles were constructed using bricks and mortars similar to those found in historic buildings in order to correlate the analytical response to the determined through laboratory testing. Ultimate load predictions were made from a finite element model in the Abaqus code, which includes the plastic behavior of the bricks and mortars. The values obtained from the tests were calibrated with numerical models. The results show that the failure modes are dependent the position of the vertical joints of the walls and the linear variation of the thickness of the walls does not correspond to a linear variation of the determined ultimate load.

1 INTRODUCTION

Most research on the behavior of unreinforced masonry walls subject to seismic actions, focus on experimental or numerical analysis [1,2,3,4,5,6]. In accordance with López J. et al [7] the masonry present a directionality in their properties because the mortar joints act as planes of weakness in the material. In general, the approach to numerical modeling can develop micromodels, which discretized bricks and joints separated. Furthermore, there is a possibility do a macro-model where the masonry be as a composite material. Significant advances have been made in the field of micro and macro modeling to identify the response of
masonry walls with different actions [8]. The analysis of masonry using the micromodel is convenient for small structural elements, with special interest in the states of stresses and strains in the panel. To represent the masonry walls behavior is needed knowledge of each element in the masonry and interaction generated between them [9]. The behavior of unreinforced masonry used in buildings subject to seismic actions can be understood through three-dimensional analyzing in panels of masonry, using computational modeling tools to help describing the behavior elastic and inelastic materials used in the constitutive model characteristics [10].

This aims of paper is to determine the structural response of unreinforced masonry piles subjected to compressive loads through a numerical simulation using a micromodel to represent the numerical model. The numerical model uses a single criterion of plasticity. The model uses information from the mechanical behavior of materials that make up the walls of unreinforced masonry. Finally the numerical results are compared with the results of the experiments.

2 CHARACTERISTICS OF THE PILES OF MASONRY MANUFACTURED IN LABORATORY

In the laboratory three piles of masonry were built. Each of the piles has a different thickness, but the other two geometric dimensions are maintained, which are equal for all specimens. Table 1 shows the geometric characteristics the piles of masonry studied.

<table>
<thead>
<tr>
<th>ID</th>
<th>Long (mm)</th>
<th>Thickness (mm)</th>
<th>Height (mm)</th>
<th>Photo</th>
</tr>
</thead>
<tbody>
<tr>
<td>PL140</td>
<td>285</td>
<td>140</td>
<td>450</td>
<td><img src="image" alt="PL140" /></td>
</tr>
<tr>
<td>PL290</td>
<td>285</td>
<td>285</td>
<td>450</td>
<td><img src="image" alt="PL290" /></td>
</tr>
<tr>
<td>PL450</td>
<td>285</td>
<td>430</td>
<td>450</td>
<td><img src="image" alt="PL450" /></td>
</tr>
</tbody>
</table>

Figure 1 show the general characteristics of the piles of masonry built in the laboratory.
Bricks are solid ceramic type were used. The dimensions of them was 290mm x 140mm and 70mm of thick. Mortar joints have an average thickness of 20mm. The mortars were prepared with a mixture in volume with the following proportions: Cement: 1; Sand: 7; common lime: 2.

Figure 2 show three mortar specimens, prepared according to the regulation CIRSOC [11]. The specimens are cubic, 70 mm side.

The mechanical behavior of the bricks and mortars was obtained from destructive testing in laboratory. Figure 3 show the compression test of mortar specimens and Figure 4 the compression tests solid ceramic bricks.
Table 2 and Table 3 show the results of the compression test for the mortars and solid ceramic bricks respectively.

**Table 2: Results of Compressive Test in Mortars.**

<table>
<thead>
<tr>
<th>ID</th>
<th>Fc [N]</th>
<th>Rc [MPa]</th>
<th>P [MPa]</th>
<th>D [MPa]</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC701</td>
<td>22900</td>
<td>4.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC702</td>
<td>21700</td>
<td>4.43</td>
<td>4.56</td>
<td>0.12</td>
</tr>
<tr>
<td>MC703</td>
<td>22370</td>
<td>4.57</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Where:
ID: Specimen identification.
Fc: Compressive load of the specimen.
Rc: Compressive stress of the specimen.
P: Average stress.
D: Dispersion.
Table 3: Results of Compressive Test in solid ceramic bricks.

<table>
<thead>
<tr>
<th>ID</th>
<th>Fc  [N]</th>
<th>σpkm' [MPa]</th>
<th>D</th>
<th>σ'pk [MPa]</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1 - 1A</td>
<td>211700</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2 - 1B</td>
<td>109200</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L3 - 2A</td>
<td>86200</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L4 - 2B</td>
<td>87000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L5 - 3A</td>
<td>109000</td>
<td>9.26</td>
<td>0.32</td>
<td>5.06</td>
</tr>
<tr>
<td>L6 - 3B</td>
<td>128700</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L7 - 4A</td>
<td>174000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L8 - 4B</td>
<td>177700</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L9 - 5A</td>
<td>151500</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L10 - 5B</td>
<td>99000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Where:
ID: Specimen identification.
Fc: Compressive load of the specimen.
σpkm': Average stress.
D: Dispersion.
σ'pk: Compressive strength [σ'pk = σpkm' (1 - 1.4xD)]

2 FEATURES NUMERICAL MODEL

With the objective of evaluate the responses of the piles of masonry subjected to vertical load, a finite element model using the Abaqus code was performed [12], (Figure 5). The model has represented the materials of the piles of masonry (bricks, mortar joints) and their respective interaction.
Concrete Damaged Plasticity model (CDP) is used to find the tension-strain relations of each piles. For modeling were used deformable solid elements in three-dimensional called C3D8R. It is a standard element of Abaqus for these problems, with 8 nodes of contact, reduced integration and hourglass control. Interaction elements were used to simulate the contact joints between brick and mortar. The contacts between surfaces were defined with the option of Abaqus "surface-surface", specifying masters surfaces (master) and slave (slave), which have the characteristic that the master surface can not be penetrated by the slave surface. The interaction of the surfaces was defined for a longitudinal and transverse behavior. For the longitudinal behavior, hard contact (Hard Contact) were specified, and for the transverse behavior a frictional behavior, setting a friction coefficient.

3.1 Concrete Damaged Plasticity Model (CDP)

Provides a general capability for modeling concrete and other quasi-brittle materials in all types of structures. Uses concepts of isotropic damaged elasticity in combination with isotropic tensile and compressive plasticity to represent the inelastic behavior of concrete. The CDP model takes into account the degradation of the elastic stiffness induced by the plastic strain in both tension and compression and adequately explains of stiffness recovery effects during cyclic load reversals.

The model requires defining the behavior in terms of stress-strain in uniaxial compression outside the elastic range. Compression stresses data are presented as a function of the inelastic strain and the stress-strain curve can be defined beyond the peak stress, in the regime of softening. Hardening data are given in terms of an inelastic strain. The compressive inelastic strain is defined as the total strain minus the elastic strain corresponding to the undamaged material [12].

Abaqus automatically converts the inelastic strain values to plastic strain values using the relationship:

\[ \varepsilon_{\text{pl}} = \varepsilon_{\text{in}} - \frac{d_{\alpha} \sigma_{\alpha} E_0}{(1 - d_{\alpha}) E_0} \]

(1)
Where:

\[ \varepsilon_{\text{pl}} = \text{plastic strain.} \]
\[ \varepsilon_{\text{in}} = \text{inelastic strain.} \]
\[ d_c = \text{stiffness degradation variable.} \]
\[ E_0 = \text{initial elasticity modulus.} \]
\[ \sigma_c = \text{compressive stress.} \]

The effective stress is defined as:

\[ \bar{\sigma} = d_0^{el} : (\varepsilon - \varepsilon_{\text{pl}}) \]  \hspace{1cm} (2)

Where:

\[ d_0^{el} = \text{initial (undamaged) elasticity matrix.} \]

The plastic flow potential function and the yield surface make use two stress invariants of the effective stress tensor, namely the hydrostatic pressure stress (eq. 3) and the Mises equivalent effective stress (eq. 4):

\[ p = -\frac{1}{3} \text{trace}(\bar{\sigma}) \]  \hspace{1cm} (3)
\[ q = \sqrt{\frac{3}{2} [S : S]} \]  \hspace{1cm} (4)

Where \( S \) is the effective stress deviator, defined as:

\[ S = \bar{\sigma} + p I \]  \hspace{1cm} (5)

3 RESULTS

Table 4 show the results obtained compression tests in laboratory to piles of the masonry.

**Table 4:** Results compressive test to piles of the masonry.

<table>
<thead>
<tr>
<th>ID</th>
<th>Weight (kg)</th>
<th>Vol (m³)</th>
<th>Specific weight (kg/m³)</th>
<th>Peak load (KN)</th>
<th>Def. in Peak load (mm)</th>
<th>Peak stress (MPa)</th>
<th>Specific Def. in Peak stress (mm/mm)</th>
<th>&quot;CMP&quot;</th>
<th>&quot;CMV&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>PL140</td>
<td>28.48</td>
<td>0.02</td>
<td>1586</td>
<td>397</td>
<td>3.43</td>
<td>9.97</td>
<td>0.0076</td>
<td>14</td>
<td>22</td>
</tr>
<tr>
<td>PL290</td>
<td>60.80</td>
<td>0.04</td>
<td>1663</td>
<td>596</td>
<td>4.22</td>
<td>7.34</td>
<td>0.0094</td>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td>PL450</td>
<td>92.40</td>
<td>0.06</td>
<td>1675</td>
<td>664</td>
<td>4.00</td>
<td>5.42</td>
<td>0.0089</td>
<td>7</td>
<td>12</td>
</tr>
</tbody>
</table>

Were:
- "CMP": Peak load divided by the pile weight.
- "CMV": Peak load divided by the pile volume.
The Table 4 show an increase of the peak load with increasing the thickness of the pile, however the opposite occurs with the peak stress, the latter decreases with increasing thickness of the pile. This effect can also be observed in Figure 6.

The results of finite element model with experimental testing was compared. Figures 7, 8 and 9 show relationships tension-strain from experimental tests and numerical analysis results.

Figure 7 show a good correlation between the numerical model and experimental testing in the elastic range, however exist differences in the plastic phase. Figure 8 show a good correlation to the PL290 pile of intermediate thickness. Figure 9 show the results of the PL450 pile. Figure 9 shows the curve of the numerical model differs from the experimental curve. Therefore the model CDP cannot efficiently capture the experimental behavior.
Figure 9 show a larger difference in the value of peak stress. The mortar volume located in the vertical joints of masonry piles can affect the response of these structures, due to the interaction generated between the brick and mortar. This phenomenon is not accurately captured by the CDP model used. It is believed that the thickness of these structures should influence how nonlinear response (in this case for stress-strain relations) situation not detected by the CDP model.

4 CONCLUSIONS
- Depending on the results of the experiments and numerical analysis concludes that a linear increase in thickness the masonry piles does not correspond to a linear increase in peak compression capacity.
- The numerical analysis tends to capture with some precision the experimental results for the thicknesses of 140 and 290 mm (PL140 and PL290). However for 450mm
thick pile (PL450), the results were significantly different.

REFERENCES


INFLUENCE OF THE PARAMETERS CHARACTERIZING THE NEAR-FAULT EARTHQUAKE IN THE SEISMIC RESPONSE OF BASE ISOLATION STRUCTURES

MIGUEL E. TORNELLO¹, CARLOS D. FRAU², DANIEL PAGANO³

¹CeReDeTeC. Facultad Regional Mendoza. Universidad Tecnológica Nacional
Rodríguez Nº 273 (5500). Mendoza. Argentina
mtornell@frm.utn.edu.ar

²CeReDeTeC. Facultad Regional Mendoza. Universidad Tecnológica Nacional
Rodríguez Nº 273 (5500). Mendoza. Argentina
cdfrau@frm.utn.edu.ar

³CeReDeTeC. Facultad Regional Mendoza. Universidad Tecnológica Nacional
Rodríguez Nº 273 (5500). Mendoza. Argentina
daniel.pagano@frm.utn.edu.ar

Key words: Seismic Protection, Seismic Isolation, Near-Fault earthquakes, Structural response.

Abstract. In the last two decades, earthquake engineering has developed new strategies to control and reduce damage to the buildings upon the occurrence of destructive earthquakes. The seismic base isolation is one of the innovative techniques of seismic protection. Ground motions near-fault, are characterized by highly destructive. Near-fault effects constructs undergo large deformations and must also dissipate considerable energy in a few cycles and reduced time. The near-fault records are characterized by velocity pulse with long period and large amplitude. The velocity pulses are those that produce the most damage on buildings. The paper evaluates the response of a three story building constructed with masonry and reinforced concrete with two types of seismic isolation devices: the elastomeric isolators with lead core and steel springs - viscous dampers. To evaluate the response of seismic isolated building, 91 near-fault records were used. Records contain different velocity pulse identified as pulse-like according to the work of Baker. The components of acceleration records were rotated at the normal and parallel to the trace of the seismogenic fault to obtain the maximum demand direction. Isolation building responses were determined by nonlinear dynamic analysis by elastomeric isolators devices and lineal analysis by springs and viscous dampers device. In both cases, the time history response was evaluated for different structural periods. The results show that there is not clear correlation between the individual parameters of near-fault ground motions and the maximum response of the isolation structure. The paper proposes a pseudo-response-spectrum, which results when the input and response are normalized by the period and the maximum velocity pulse amplitude respectively.
1 INTRODUCTION

Ground motions close to a fault rupture can be significantly more different than those farther away from the seismic source. The near-fault zone is typically assumed to be within a distance of about 15-20 km from a fault rupture. Within this near-fault zone, ground motions are significantly influenced by the rupture mechanism, the direction of rupture propagation relative to the site, and possible permanent ground displacements resulting from the fault slip [1]. Ground Motions near-fault is typically characterized by a motion pulse-type of short duration and large amplitude. This motion concentrates the input energy to the structures in a few pulses at the beginning of the record. Structures placed in near-fault zones need special considerations in the seismic design [2].

In the near-fault region, structural damage occurs for one or two severe cycles of inelastic deformation. These cycles correspond to ground motions with long and large pulses of velocity. Ground motions with long and large pulses of velocity can generate a much higher base shear, inter-story drift, and roof displacements in high-rise buildings compared with ground motions that do not contain these pulses.

The ductility demand can also be much higher and the effectiveness of supplemental damping can be much lower both for pulse-like ground motions [3]. In high-rise buildings, ground motions with large pulses of velocity and displacement cause in the structural response a large participation of higher modes [4]. There exist numerous studies and applications related to the structural response with seismic isolation, nevertheless, few researcher bear in mind seismic sources characteristics [5].

The displacement of the isolated structures subject to near-fault ground motion is strongly influenced by one of the ground motions components [6]. Investigations in structures with natural rubber bearing isolators subject to near-fault ground motion indicate that an increase in the damping of isolation devices achieve minor displacements, inter-story drift, seismic base shear, accelerations and velocity [7]. Investigations realized by Naeim F. et al. [8] indicated that increasing the damping of the isolation device, reduces the displacement but increases the accelerations and inter-story drift. Nevertheless, there is no indication of the seismic parameter controlling the structural response when the record possesses long pulses of velocity and displacement or how to control the dimensions of the isolations system before the presence of the mentioned pulses.

Numerous analysis and design procedures for structures with passive energy dissipation systems are present in specialized literature [9]. There are documents and standards which establish requirements for such structures [10]. But there are not many developments about the dissipation devices requirements in structures which could be submitted to near-fault ground motion.

Earthquakes are complex actions on structures and cannot be represented by a single parameter. The action of earthquakes is often characterized by a single parameter (PGA) or by the elastic spectral response for 5% of damper and for a single period. In other cases the spectrum is modified by a factor that takes into account the duration of the earthquake (HAZUS methodology). However these standard methods underestimate the damage in structures where a single parameter of earthquake is employed [11].

A structure may have more damage to earthquakes more duration respect to earthquake a few seconds duration if the PGA earthquakes of both earthquakes are the same [11]. Seyedi
work shows that when in evaluating the responses of structures, two or three parameters that characterize the ground motion, a significant reduction in the dispersion curve fragility is obtained. A strong correlation between the spectral ordered to 1s and the maximum value of the story drift is obtained. Some parameters such as the PGA and Arias intensity have an important influence on the maximum value of story drift. Other parameters such as duration or number of cycles are not useful for predicting the damage to the building [11].

The aim of this paper is to evaluate the response of two types of seismic isolation devices subjected to near-fault ground motions. The characteristics of the near-fault ground motion are presented. A set of this type of ground motion is selected for numerical analysis. We considered one structures with two type of seismic isolation: i) hybrid isolation system composed by steel springs and viscous dampers (GCS), ii) lead core elastomeric bearings (LRB). Non linear time history analyses were carried out. Finally the incidence of ground motion parameters in the response of structures with base isolation for both devices is analyzed.

2 STRUCTURE WITH SEISMIC ISOLATION DEVICE

This building possesses three levels with concrete structure, masonry walls and concrete slab. Plant dimensions are 8.00 x 7.60 m. When a participation of 25% of the live load is assumed, the weight of the building is 2570 KN and 2910 KN when participation the live load is 100%. The building period is 0.89 s with seismic isolation and 0.17 s, for the same building, but with fixed base [12] (Figure 1.a). Seismic isolation device consist of four steel spring packages (GCS, GERB® Control Systems) and viscous elastic dampers with vertical axis (GerbVisco®) (Figure 1.b). The devices installed correspond to the model EQ-07 with a vertical load capacity of 921 KN, a vertical stiffness of 35.40 KN/mm and a horizontal stiffness of 4.73 KN/mm. The damping design was 26 % in horizontal direction and 13 % in vertical direction. A model in finite elements in 3D was used in the design of the building with seismic isolation (Figure 1.b). Damping force–velocity ratio of the viscous damper is shown in Figure 2.a.

![Figure 1.a: Building with seismic isolation.](image)

![Figure 1.b: Steel spring packages and viscous dampers (GCS)](image)

To quantify the influence of the parameters that characterize earthquakes in the response of isolated building with GCS devices, results for periods of Tn=1.57s and Tn=2.03 s were analyzed too.
The second device used to evaluate the response of the building consists of four lead core elastomeric bearings (LRB) located at each corner of the building. To verify the design parameters Chilean regulation [13] was used. The diameter of LRB is 600 mm, the diameter of lead core is 150 mm and the total height was 268 mm. The effective stiffness is 4241.70 KN/m with vertical stiffness of 5437185 KN/m and 17.60% effective damping. The building period is 1.44 s with LRB device. To quantify the influence of the parameters that characterize earthquakes in the response of isolated building with LRB devices, results for periods of $T_n=1.21$ and $T_n=2.52$ s were analyzed too. The constitutive curve of the isolators that was used for the non-linear dynamic model is shown in Figure 3 and Table 1.

![Graph](image)

**Figure 2.a:** Damping force-velocity ratio  
**Figure 2.b:** Element finite model in 3D

**Figure 3:** Constitutive characteristic of LRB isolator.

<table>
<thead>
<tr>
<th>Table 1: Dimension of LRB isolator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isolator’s characteristic</td>
</tr>
<tr>
<td>Isolator’s diameter</td>
</tr>
<tr>
<td>Lead core diameter</td>
</tr>
<tr>
<td>Total Height</td>
</tr>
<tr>
<td>Plate rubber thickness</td>
</tr>
<tr>
<td>Plate steel thickness</td>
</tr>
</tbody>
</table>
Structural response by time history nonlinear dynamic analysis is obtained. The software used for such target was SAP2000 [14]. The analysis is based on the proper of the viscous linear damping and not proportionality between the stiffness and mass. Nonlinearity is concentrated on the isolation devices and the building is considered as a rigid block and elastic. In this case the unbalanced non lineal force in every time step, are analyzed by mean of a number of reduced structural modes [15]. The method of direct integration of the equilibrium equations represents appropriately the behavior of the seismic isolation but only it allows analyzing deterministic sign in the time dominion. Preliminary studies [12] compared the structural response obtained in theoretical form between the building with seismic isolation and another with fixed base of identical characteristics.

3 GROUND MOTION NEAR-FAULT SELECTED

Several parameters have been used to characterize ground motion. The most familiar parameters are PGA, PGV and PGD (peak ground acceleration, velocity and displacement). Pulse-like motions have been identified as critical in structures design in the near-fault zone. The analysis of elastic and inelastic multiple degree of freedom systems indicates that the amplitude and period of pulse in the velocity-time history (Av=PGV in these cases) and Tv are parameters that control the performance of structures. Lara et al [16] demonstrated that the Maximum Variation of Ground Velocity (largest peak to peak value in the ground velocity) is an important cause of inelastic response for some structures. Malhotra [3] showed that near-fault ground motions with directivity effects tend to have high PGV/PGA ratio and this ratio dramatically influences response characteristics. To evaluate the response of seismic isolated building 91 near-field records were used, the records identified as pulse-like according to the work of Baker [17]. The components of acceleration records were rotated at the normal and parallel to the trace of the seismogenic fault to obtain the maximum demand direction. In the processing of 91 seismic records were identified: pulse period (Tv), peak velocity (PGV) and maximum values of acceleration (PGA).

3.1 Methodology and isolation building response

To evaluate the response of isolated building with two isolation systems and three periods, 91 selected seismic records were analyzed. To find the response of the structure, time history analysis was made; lineal for the GCS device and nonlinear for LRB devices. Were determined for the level of slab above the isolation devices and the last level of the isolation building the following parameters: maximum displacement, relative velocity, absolute velocity and relative and absolute acceleration; finally the maximum and minimum basal shear was calculated.

4 RESULTS ANALYSIS

In order to find correlations between the parameters of earthquakes and the structural response, various combinations were studied. As characteristic parameters of earthquakes were considered the PGA, PGV and the period of greater velocity pulse Tv. Parameters used for evaluating the structural response in structures whith seismic isolation were a normalized base shear $|V|=V_i/V_{\text{max}}$, where “$V_i$” value maximum response for ground motion “$i$”, and “$V_{\text{max}}$” greater response value for all ground motions. It is assumed that the isolation devices...
are able to absorb the resulting displacement.

4.1 Results GCS isolation device

For each period studied (0.89 s, 1.57 s and 2.03 s) and for the 91 ground motion selected, structural responses versus parameters of earthquakes (PGA, PGV, PGV / PGA and Tv), were plotted. In Figure 4 some of the graphs obtained are shown. In order to compare the different graphs, the vertical axis, representing the structural response, has been normalized by dividing the maximum value found for the set of records considered |V| = Vi / Vmax.

![Figure 4: Structural responses versus parameters of earthquakes for GCS device](image)

The above relations, presents a significant dispersion in the results and no clear impact of the individual parameters of the seismic record in the structural response, for example, in terms of the shear base. The above figures indicate that there is not a maximum response of isolation building, for a characteristic parameter of earthquake.

4.2 Results LRB isolation device

For each period studied (1.21 s, 1.44 s and 2.52 s) and for the 91 ground motion selected, structural responses versus parameters of earthquakes (PGA, PGV, PGV / PGA and Tv), were plotted. In Figure 5 some of the graphs obtained are shown. In order to compare the different graphs, the axis representing the structural response (vertical axis) has been normalized by dividing the maximum value found for the set of records considered |V| = Vi / Vmax.
dividing the maximum value found for the set of records considered $|V| = \frac{V_i}{V_{\text{max}}}$.

The above relations, presents a significant dispersion in the results and no clear impact of the individual parameters of the seismic record in the structural response, for example, in terms of the shear base. The above figures indicate that there is not a maximum response of isolation building, for a characteristic parameter of earthquake.

4.3 Pseudo-Response-Spectrum (PRS)

In order to find a correlation between the characteristic parameters of the earthquake and the response of the structure, the results obtained are represented as follows. On the horizontal axis the ratio $T_n/T_v$ is represented, where $T_n$ is the period of the structure and $T_v$ is the period of greatest velocity pulse in the seismic record. On the vertical axis is plotted the maximum response divided by the PGV ($V_i/\text{PGV}$). These representations are shown in Figure 6.
In Figures 6 on points, was manually traced an outline to the maximum responses. This contour takes the form of an acceleration response spectrum. For each period, the maximum response is always found for the ratio \( T_n/T_v \) approximately equal to one. It is observed that, for the relation \( T_n/T_v=1 \), the maximum value of \( \frac{V_i}{PGV} \) decreases as increasing the period of building isolated.

5 CONCLUSION

The structural response of the building with devices isolates studied (GCS and LRB device) do not show a clear correlation with the individual parameters of the near-fault
ground motions such as PGA, PGV and Tv; neither no clear correlation to the ratio (PGV / PGA) appears. However, it is observed that the PGV and Tv are parameters that affect the response.

To find a relationship between the structural response and the parameters that characterize the near-fault ground motions, authors propose working with normalized parameters. On the vertical axis, that represents the maximum response (V), is normalized by the maximum peak of velocity pulse (PGV). The horizontal axis, that represents each ground motion, is normalized by dividing the isolation building period (Tn) and the period of the maximum velocity pulse (Tv).

Following the normalization indicated above, the results appear with a form of acceleration response spectrum. Thus, the paper proposes a Pseudo-Response-Spectrum (PRS) in which the maximum response is obtained for the ratio (Tn / Tv) = 1, so that the period of isolation building coincides with the period of maximum pulse of seismic record. This feature is independent of the structural period. In increasing the period of isolation building, a reduction of the structural response is observed. This result is not affected by the type of isolation device studied.

The PRS proposed could be helpful to estimate the maximum response of isolation building located in regions near-fault. With the period of the isolation structure and the characteristics of higher velocity pulse expected in the building site, it is possible to determine the maximum response in terms basal shear.

ACKNOWLEDGEMENTS

This work was developed within the project research 25-J091 funded by the National Technological University. Collaboration of engineer Emilce Giolo, engineer Maria Soledad Garcia and student Alejandro Forte for further information processing work is appreciated.

REFERENCES


PERFORMANCE-BASED DESIGN OPTIMIZATION OF STRUCTURES UNDER SEISMIC DEMANDS, INCORPORATING PASSIVE CONTROL DEVICES

OSCAR MÖLLER¹, RICARDO O. FOSCHI² AND JUAN P. ASCHERI¹

¹Instituto de Mecánica Aplicada y Estructuras (IMAE), Facultad de Cs. Exactas, Ingeniería y Agrimensura, Universidad Nacional de Rosario
Riobamba y Berutti, 2000 Rosario, Argentina
moller@fceia.unr.edu.ar

²Civil Engineering Department, University of British Columbia
6250 Applied Sciences Lane, Vancouver, B.C., Canada V6T 1Z4
rowfa1@civil.ubc.ca

Key words: Performance-based seismic design, Optimization, Control devices, Reliability, Social costs.

Abstract. Performance-based design in earthquake engineering is a structural optimization problem that has, as the objective, the determination of design parameters for the minimization of total costs, while at the same time satisfying minimum reliability levels for the specified performance criteria. Total costs include those for construction and structural damage repairs, those associated with non-structural components and the social costs of economic losses, injuries and fatalities. This paper studies the influence, on the optimization results, of incorporating energy dissipation devices based on steel yielding. We use a general framework to approach this problem, using a numerical optimization strategy and incorporating the use of neural networks for the evaluation of dynamic responses and the reliability levels achieved for a given set of design parameters. The strategy is applied to an example of a four-story office building, with energy dissipation devices located at each floor level. Optimum results show that a minimum control force and stiffness is necessary to reduce the replacement of control devices after the occurrence of earthquakes, which increases the total life cycle cost (LCC).

1 INTRODUCTION

The general objective of performance-based design of structures, under different seismic demands, is to obtain values for several design parameters which result in the satisfaction of performance requirements with specified minimum reliabilities while, at the same time, achieving a minimum for the objective function.

Here, this function is taken to be the total cost over the service life (LCC). This total cost includes: (i) the initial construction cost of the resisting structure, of the non-structural elements and the building contents; (ii) repair costs associated with the occurrence of earthquakes during the building’s service life; and (iii), social costs, including those associated with injuries and death in case of severe damage and structural collapse, plus costs
related to loss of business and stocks.

This work considers, in particular, the influence in the optimization of incorporating energy dissipation devices, based on steel yielding. These devices reduce considerably the structural displacements and the corresponding damage, but must be replaced when, after an earthquake, they have undergone plastic deformation. The initial cost of the devices, and their replacement, must also form part of the life cycle cost.

The problem considered is that of seismic design for a reinforced concrete structure, for an office building, including energy dissipation devices at each floor level. The solution must consider all the uncertainties present, either related to the ground motions or to the structural behaviour. In the context of performance-based seismic design [1,2], performance or limit states functions are specified for different performance levels: operational, life safety, and collapse, with acceptable or target minimum reliabilities specified for each level.

The initial step in the solution is the determination of the structural response parameters, maxima during the earthquake (maxima displacements, inter-story drifts, accelerations, ductility demand in the energy dissipation devices). These values result from nonlinear dynamic analyses, for each specific combination of the variables in the problem and each ground motion. These responses are discrete values and, for estimation of probabilities of failure, it is convenient to represent them with continuous functions. These approximations, called response surfaces, can then be used as substitutes for the true structural responses for combinations not considered [3]. Different forms of response surfaces have been studied [4], with neural networks offering advantages of flexibility and adaptability, and have been adopted in this work. Estimating responses via a response surface, whichever form is adopted, facilitates the evaluation of the performance functions and makes probability estimation via Monte Carlo simulation very efficient [5].

Here we consider the optimization of structural parameters like element dimensions and reinforcement ratios, while maintaining fixed the layout of the building. The optimization strategy must consider minimum reliability constraints, the dimension, form and number of performance functions. Optimization algorithms can implement different strategies [6,7,8], some requiring the calculation of gradients within schemes of steepest descent or of conjugate gradients [9,10]. Other strategies are not gradient-dependent, and utilize heuristic schemes like random search or genetic algorithms [11,12,13].

Here we utilize an algorithm for a random search that has been developed in previous work [14,15,16]. Starting from an initial design, the optimization parameters are randomly perturbed within a search zone. For each choice of parameters, satisfaction of reliability constraints is verified, and the total cost is evaluated. The combination corresponding to the local minimum is chosen as the anchor combination for the next search zone. The search is continued until convergence is achieved, in the sense that no total cost within the current search zone is found to be less than that for the anchor. In the applications described here, the optimization allows an evaluation of the importance of the cost of the energy dissipation devices, as well as obtaining interesting results in reference to the failure probabilities at the optimum solution.
2 STRUCTURE TO BE OPTIMIZED, INCLUDING CONTROL DEVICES

The structure to be optimized is a four-story office building with spans and columns as shown in Figure 1, both in plan and for a portal along the x-direction, which is the assumed direction of the earthquake motion.

Figure 1: Structural system
3 THE GENERAL PROCESS OF OPTIMIZATION

Figure 2 shows the general optimization process, which is divided into blocks that can be executed separately but in sequence.

**Block 1: STRUCTURAL ANALYSIS**
Definition of the intervening variables $X$ with their bounds
Experimental design: choose a combination $x_k$ from the set $X$
Nonlinear dynamic analysis to obtain responses $R_i(x_k)$ for $x_k$
Response approximation via a neural network, $F_i(X)$

**Block 2: RELIABILITIES**
Definition of the performance functions $→ G_j(X)$
Definition of the design parameters $x_d$ with their bounds
Experimental design: choose a combination $x_{dk}$ from the set $x_d$
Reliability $j$ associated with the combination $x_{dk} → \beta_j(x_{dk})$
Reliability approximation via neural networks $\beta_j(x_d)$

**Block 3: OPTIMIZATION**
Definition of the objective function $→ C(x_d)$
Definition of target minimum reliability constraints $→ \beta_{JT}$
Preliminary design: $x_{d0}$
Optimization algorithm, to obtain $x_d$ for minimum $C(x_d)$, with $\beta_j(x_d) ≥ \beta_{JT}$

**Figure 2:** Block organization for the optimization process

3.1 Structural analysis

Table 1 shows the variables considered in the calculation of the discrete responses, along with the lower and upper bounds. The reinforcement ratios have bounds corresponding to requirements from building Codes, in order to provide minimum strength against gravitational loads. The design procedure [17] is based on the load combination $1.0 D + 0.25 L + E$, which, for the structure studied here, combines a permanent load $D = 7.66 KN/m^2$ and a live load $L = 3 KN/m^2$, giving a total weight per floor of $W = 1817 KN$. Because the slabs are considered
rigid in their plane, each portal frame must support 1/5 of the total seismic load, resulting in a mean value for the mass of the beams \( m = 3.09 \times 10^{-4} \text{ KN s}^2 / \text{cm}^2 \).

The seismic action \( E \) corresponds to ground motions with accelerograms artificially generated [18,19], for which the main variables are the peak ground acceleration \( a_G \) and the central frequency for the ground filter, \( f_g \). Other variables used in the generation of accelerograms are either functions of \( a_G \) and \( f_g \), or are taken as constants.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Lower bound</th>
<th>Upper bound</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X(1) = m ) (KN s(^2) / cm(^2))</td>
<td>2.00\times10^{-4}</td>
<td>4.00\times10^{-4}</td>
<td>Beam mass per unit length</td>
</tr>
<tr>
<td>( X(2) = b_b ) (cm)</td>
<td>15</td>
<td>30</td>
<td>Beam section width</td>
</tr>
<tr>
<td>( X(3) = h_b ) (cm)</td>
<td>30</td>
<td>70</td>
<td>Beam section depth</td>
</tr>
<tr>
<td>( X(4) = b_c ) (cm)</td>
<td>20</td>
<td>40</td>
<td>Column section width</td>
</tr>
<tr>
<td>( X(5) = h_c ) (cm)</td>
<td>30</td>
<td>100</td>
<td>Column section depth</td>
</tr>
<tr>
<td>( X(6) = \rho_b )</td>
<td>0.00298</td>
<td>0.01389</td>
<td>Beam reinforcement ratio (midspan)</td>
</tr>
<tr>
<td>( X(7) = \rho'_b )</td>
<td>0.00298</td>
<td>0.01389</td>
<td>Beam reinforcement ratio (supports)</td>
</tr>
<tr>
<td>( X(8) = \rho_c )</td>
<td>0.008</td>
<td>0.04286</td>
<td>Column reinforcement ratio</td>
</tr>
<tr>
<td>( X(9) = f_r / f_{c0} )</td>
<td>0</td>
<td>0.15</td>
<td>Confinement pressure (normalized)</td>
</tr>
<tr>
<td>( X(10) = F_y ) (KN)</td>
<td>0</td>
<td>2850</td>
<td>Yield force of control device</td>
</tr>
<tr>
<td>( X(11) = K_d ) (KN / cm)</td>
<td>0</td>
<td>49128</td>
<td>Stiffness of control device</td>
</tr>
<tr>
<td>( X(12) = a_G ) (cm / s(^2))</td>
<td>10</td>
<td>1200</td>
<td>Peak ground acceleration</td>
</tr>
<tr>
<td>( X(13) = f_g ) (Hz)</td>
<td>1.50</td>
<td>3.50</td>
<td>Central ground filter frequency</td>
</tr>
</tbody>
</table>

Within the bounds for the variables, and using techniques of experimental design, \( N = 648 \) variable combinations are chosen randomly, adequately covering the entire domain. For each combination, \( NS = 10 \) sub-combinations are generated taking into account the following secondary random variables: (a) A set of random phase angles applied to the different frequencies in the accelerogram, to be scaled using the peak acceleration \( a_G \) included in the main variable combination as \( X(12) \); (b) The concrete strength and that for the steel reinforcement, which generate the variability in the momento-curvature relationships for both the beams and the columns.

For each of the ten sub-combinations associated with one main combination, a nonlinear dynamic analysis is carried out for the structure of Figure 1, utilizing a model with bar elements [19,20]. Each analysis provides the following responses entering into the Limit States: \( UMAX \), horizontal maximum displacement of the top floor; \( AMAX(i) \), \( i = 1,2,3 \): maximum acceleration for, respectively, the first, the second and the third floor; \( DISTM: \)
maximum inter-story drift; \(DIES\): global damage index for the structure; \(DILOM\): maximum local damage index; \(DUCD(i), i = 1,2,3,4\): maximum ductility demand in the dissipation devices for each floor level.

Being \(R\) a generic response (e.g., \(DISTM\)), for each main combination \(j\) and subcombination \(k\), the responses \(R_{kj}\), \((j = 1, NP, k = 1, NS)\) are used to calculate the mean value and standard deviation of the response over the set of sub-combinations:

\[
\overline{R}_j = \frac{1}{NS} \sum_{k=1}^{NS} R_{kj} \quad \sigma_{R_j} = \sqrt{\frac{1}{NS-1} \sum_{k=1}^{NS} (R_{kj} - \overline{R}_j)^2}
\]  

(1)

The \(NP\) discrete results \(R(X)\) from (1) are then represented by two response surfaces: one for the mean values and another for the standard deviations. Here we utilize a neural network \(F(X)\) as the response surface,

\[
R(X) \approx F(X) = h \left( \sum_{j=1}^{J} W_{kj} h(\sum_{i=1}^{N} W_{ji} X_i + W_{j0} + W_{k0}) \right)
\]  

(2)

in which \(R(X)\) is the “true” response from the dynamic analysis for the variables \(X_i\), \(F(X)\) is the neural network approximation, \(W_{kj}\) and \(W_{ji}\) are constant weights to be determined, and \(h(t)\) is a function transferring information from one layer of the neural network to the next. A neural network training or calibration consists on determining the weights \(W\) so that the differences between predictions \(F(X)\) and the data \(R(X)\) are minimized.

The approximation of \(F(X)\) with \(R(X)\) is never perfect, and the scattering of differences can be quantified using the standard deviation of the relative error, which is an average coefficient of variation, as follows:

\[
\sigma_{\varepsilon_r} = \sqrt{\frac{1}{NP-1} \sum_{k=1}^{NP} \left( \frac{Y_k - T_k}{Y_k} \right)^2}
\]  

(3)

in which \(Y_k\) is the value calculated with the network, \(T_k\) is the value obtained with the dynamic analysis and \(NP\) is the number of combinations used in the database.

The mean value and the standard deviation of the responses over the sub-combinations can then be written taking into account the dispersion error:

\[
\overline{F}(X) = \overline{Y}(X) \left( 1 + \sigma_{\varepsilon_m} X_{N_1} \right) \quad \sigma_f(X) = \sigma_y(X) \left( 1 + \sigma_{\varepsilon_m} X_{N_2} \right)
\]  

(4)

in which \(\overline{Y}(X)\), \(\sigma_y(X)\) are the mean value and standard deviation of the response as calculated from the corresponding neural network, and \(X_{N_1}\), \(X_{N_2}\) are Standard Normal random variables. Finally, using a Lognormal distribution for the response \(R(X)\) over the sub-combinations, \(R(X)\) can be written,

\[
R(X) \approx F(X) = \exp \left[ \overline{F}(X) \sqrt{\ln(1 + \frac{(\sigma_f(X))^2}{F(X)})} \right]
\]  

(5)
3.2 Reliabilities

Table 2 shows the random variables, along with their probability distribution and statistics. The sign ? indicates a design parameter in the optimization, that is, the mean values for the dimensions of beams and columns, the reinforcement ratios and the strength parameters for the energy dissipation devices.

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\bar{X}$</th>
<th>$\sigma_X$</th>
<th>Type</th>
<th>Variable</th>
<th>$\bar{X}$</th>
<th>$\sigma_X$</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X(1) = m$</td>
<td>$3.09 \times 10^{-4}$</td>
<td>$3.09 \times 10^{-5}$</td>
<td>Normal</td>
<td>$X(11) = K_d$</td>
<td>?</td>
<td>$0.10 \bar{X}$</td>
<td>Lognormal</td>
</tr>
<tr>
<td>$X(2) = b_b$</td>
<td>$20 \text{ cm}$</td>
<td>$1 \text{ cm}$</td>
<td>Normal</td>
<td>$X(12) = a_G$</td>
<td>$48 \text{ cm/s}^2$</td>
<td>$78 \text{ cm/s}^2$</td>
<td>Lognormal</td>
</tr>
<tr>
<td>$X(3) = h_b$</td>
<td>? cm</td>
<td>$0.05 \bar{X}$</td>
<td>Normal</td>
<td>$X(13) = f_e$</td>
<td>$2.50 \text{ Hz}$</td>
<td>$0.375 \text{ Hz}$</td>
<td>Normal</td>
</tr>
<tr>
<td>$X(4) = b_c$</td>
<td>$30 \text{ cm}$</td>
<td>$1.5 \text{ cm}$</td>
<td>Normal</td>
<td>$X(14) = \sigma_{\pi_G}$</td>
<td>0</td>
<td>0.25</td>
<td>Normal</td>
</tr>
<tr>
<td>$X(5) = h_c$</td>
<td>? cm</td>
<td>$0.05 \bar{X}$</td>
<td>Normal</td>
<td>$X(15) = a_G$</td>
<td>$X(15) = X(12) [1.0+X(14)]$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X(6) = \rho_s$</td>
<td>?</td>
<td>$0.10 \bar{X}$</td>
<td>Lognormal</td>
<td>$X(16) = R_{N1}$</td>
<td>0</td>
<td>1</td>
<td>Normal</td>
</tr>
<tr>
<td>$X(7) = \rho'_s$</td>
<td>?</td>
<td>$0.10 \bar{X}$</td>
<td>Lognormal</td>
<td>$X(17) = X_{N1}$</td>
<td>0</td>
<td>1</td>
<td>Normal</td>
</tr>
<tr>
<td>$X(8) = \rho_{col}$</td>
<td>?</td>
<td>$0.10 \bar{X}$</td>
<td>Lognormal</td>
<td>$X(18) = X_{N2}$</td>
<td>0</td>
<td>1</td>
<td>Normal</td>
</tr>
<tr>
<td>$X(9) = f_r / f'_c$</td>
<td>0.10</td>
<td>0.01</td>
<td>Normal</td>
<td>$X(19) = X_{N3}$</td>
<td>0</td>
<td>1</td>
<td>Normal</td>
</tr>
<tr>
<td>$X(10) = F_y$</td>
<td>?</td>
<td>$0.10 \bar{X}$</td>
<td>Lognormal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

It is assumed that the ordinate of the INPRES-CIRSOC 103 Parte I [17] design spectrum has an exceedence probability of 10% in 50 years, and for the zone that includes the city of Mendoza the mean peak ground acceleration is $a_G \approx 0.35 \text{ g}$. Following the procedure from FEMA 356 [2], accelerations corresponding to other exceedence probabilities can be obtained, and this set can then be represented by a Lognormal distribution for $a_G$ with a mean of $48.0 \text{ cm/s}^2$ and a standard deviation of $78 \text{ cm/s}^2$.

Eqs.(8) to (14) describe the nine functions adopted to describe the structural performance at three different levels: operational, life safety and collapse. A random variable $X_{N3}$, Standard Normal, is introduced in order to take into account modelling error in the calculation of the demand parameters. It is assumed that these quantities show an uncertainty with a coefficient of variation $COV = 0.10$.

- **Operational:**
  \[ G_{11}(X) = 0.005 - DISTM(X) \left[ 1 + COV \cdot X_{N3} \right] \quad \text{mode 1} \]  \hfill (6)
  \[ G_{12}(X) = 0.10 - DILOM(X) \left[ 1 + COV \cdot X_{N3} \right] \quad \text{mode 2} \]  \hfill (7)
  \[ G_{13}(X) = 1.00 - DUCDM(X) \left[ 1 + COV \cdot X_{N3} \right] \quad \text{mode 3} \]  \hfill (8)

- **Life Safety:**
  \[ G_{21}(X) = 0.015 - DISTM(X) \left[ 1 + COV \cdot X_{N3} \right] \quad \text{mode 4} \]  \hfill (9)
\[ G_{22}(X) = 0.40 - DIES(X) \left[ 1 + COV \ X_{N_3} \right] \] \quad \text{mode 5} \tag{10} \\
\[ G_{23}(X) = 0.60 - DILOM(X) \left[ 1 + COV \ X_{N_3} \right] \] \quad \text{mode 7} \tag{11} \\
- \text{Collapse} \\
\[ G_{31}(X) = 0.025 - DISTM(X) \left[ 1 + COV \ X_{N_3} \right] \] \quad \text{mode 6} \tag{12} \\
\[ G_{32}(X) = 0.80 - DIES(X) \left[ 1 + COV \ X_{N_3} \right] \] \quad \text{mode 7} \tag{13} \\
\[ G_{33}(X) = 1.00 - DILOM(X) \left[ 1 + COV \ X_{N_3} \right] \] \quad \text{mode 7} \tag{14} \\

Within the bounds of the design parameters, and applying again experimental design, \( MC = 335 \) combinations are chosen at random for the design parameters \( x_d \). For each combination, a Monte Carlo simulation is used to determine the failure probability (or reliability index) for each of the 7 limit states or failure modes in Equations (6) to (14). The discrete results \( \beta_j(x_d) \), \( j=1,7 \) are then represented by neural networks, which are utilized during the optimization to verify the compliance with the imposed minimum reliability constraints.

### 3.3 Optimization

The objective function for the optimization is the total cost. This includes the initial cost of construction, \( C_0(x_d) \), plus the repair costs \( C_d(x_d) \) and the social costs \( C_s(x_d) \) due to the occurrence of earthquakes during the service life of the structure. Accordingly,

\[ C(x_d) = C_0(x_d) + C_d(x_d) + C_s(x_d) \tag{15} \]

The initial cost of construction includes that for beams and columns, \( C_{01}(x_d) \), plus the cost of the energy dissipation devices and their attachments, \( C_{02}(x_d) \). Costs associated with slabs and foundations do not depend on the design parameters, and are taken here as a constant \( C_{03} = $1404000 \). Similarly, the cost of non-structural elements is taken as \( C_{04} = $5120000 \), and the cost of contents (equipment and furnishings) as \( C_{05} = $2265000 \).

The repair costs, at present values, depend on the level of damage caused by the earthquakes, the uncertainty associated with their arrival, the number of earthquakes during the life \( T_D \) of the structure and the interest rate \( r = 0.05 \) available for a repair fund from the time of construction until the occurrence of the damages. If \( PR \) is a response used to quantify the damage to the structure, the non-structure elements and the contents, and if \( C_d(PR) \) is the cost of repairs required at a time \( t \), under the assumption that the structure is repaired after each event, returning it to the original conditions, the expected cost \( C_d|_{PR} \) (at present values and conditional on the response \( PR \)) becomes [16]:

\[
C_d|_{PR} = \sum_{n=1}^{\infty} C_f(PR) \nu \sum_{i=0}^{n-1} \frac{\nu^i}{i!} \int_0^{T_D} t^i \exp(-(r+\nu)t) \, dt \cdot \left( \frac{\nu T_D}{n!} \right) \exp(-\nu T_D) \tag{16}
\]

in which \( \nu = 0.20 \) is the mean arrival rate of the earthquakes for Mendoza city, and \( n \) is the number of earthquake events in \( T_D \). Finally, the total repair cost is calculated by integration of the conditional probability \( C_d|_{PR} \) with the density function for the damage response parameter \( PR \). For damage to the structure we use \( PR = DIES \); for damage to the energy dissipation devices we use \( PR = DUCDM \); for damage to non-structural elements we use the inter-story drift, \( PR = DISTM \); and, finally, for damage to the contents we use \( PR = ACELM \).
Social costs, which must be assumed by the businesses operating in the building, include the costs of re-insertion into the work routine, medical costs and recovery in the case of injuries, compensation payments in the case of loss of life, losses due to business interruption and re-stocking, and rental costs during re-construction. These costs are conditional on the damage parameter $DIES$, and the total must be obtained by integration using the probability density function for $DIES$. Thus, the social costs are function of the design parameters $x_d$ through the dependence of $DIES$ on those parameters.

The algorithm used for optimization is based on a random search, without calculation of gradients. The procedure, [14], within the domain of $x_d$, starts from an initial choice for the design parameters, or first “anchor point”, $x_{d0}$. A set of $n$ combinations of $x_d$ are randomly chosen in the vicinity of $x_{d0}$, within a “search zone”. The reliability constraints are verified for each combination, and should any constraint be violated, a new combination is chosen. The total cost is calculated for each combination, and if the minimum cost is lower than that for the anchor point, the corresponding combination becomes the new anchor point and the process is repeated. The process stops when none of the $n$ combinations within the search zone has a total cost lower than the one for the anchor. This anchor provides, then, an approximation to the optimum solution.

The target maximum annual failure probabilities $P_{f\text{annual}}$ are chosen according to the recommendations by Paulay y Priestley [21]: $2 \times 10^{-2}$ for operational performance; $2 \times 10^{-3}$ for life safety and $2 \times 10^{-4}$ for the limit state of collapse. Considering that earthquakes for the city of Mendoza obey a Poisson arrival process with a mean rate of $\nu = 0.20$, those annual probability limits are equivalent to the following reliability indices for the event of earthquake occurrence: $1.276$ (operational), $2.326$ (life safety), $3.090$ (collapse).

### 4 NUMERICAL RESULTS AND DISCUSSION

Table 3 presents a summary of the optimization results: in two cases, 01 and 02, the reliability constraints from Section 3 are enforced, and the two cases correspond to two replications starting from different initial conditions; in the other two cases, 03 and 04, the reliability constraints are not enforced.

For cases 01 and 02, when reliability constraints are imposed, the optimum design is controlled by the reliability index $\beta_3(x_d)$ corresponding to yielding of the dissipation devices. For $\beta_3(x_d) \geq \beta_{3T} = 1.273$, all other reliability indices amply exceed their minima. The structural dimensions are quite similar, with corresponding similar initial costs.

More differences are observed for the repair costs of the passive control devices, being greater in case 01 than in case 02. This difference is explained by observing that the dimensions for columns and devices are slightly smaller in case 01, which produce lower reliability indices $\beta_4(x_d)$ and $\beta_6(x_d)$, corresponding to the limit states of inter-story drift for the performance levels of life safety and collapse. As more inter-story drift is allowed, more is the ductility demand in the control devices and higher is their repair cost.

It can also be observed that the repair costs for damage to structural elements, $C_{d1}$, and to non-structural elements, $C_{d3}$, are small or negligible. This indicates that the dissipation of energy is concentrated in the passive control devices, which are replaceable, thus protecting the other components in the building.
<table>
<thead>
<tr>
<th>Results</th>
<th>With reliability constraints</th>
<th>Without reliability constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_d(1) = \bar{X}(3) = h_b$ [cm]</td>
<td>55.0</td>
<td>50.0</td>
</tr>
<tr>
<td>$x_d(2) = \bar{X}(5) = h_c$ [cm]</td>
<td>52.9</td>
<td>55.8</td>
</tr>
<tr>
<td>$x_d(3) = \bar{X}(6) = \rho_b$</td>
<td>0.01057</td>
<td>0.01137</td>
</tr>
<tr>
<td>$x_d(4) = \bar{X}(7) = \rho'_b$</td>
<td>0.01019</td>
<td>0.00876</td>
</tr>
<tr>
<td>$x_d(5) = \bar{X}(8) = \rho_c$</td>
<td>0.02775</td>
<td>0.02741</td>
</tr>
<tr>
<td>$x_d(6) = \bar{X}(10) = F_y$ [KN]</td>
<td>426</td>
<td>477</td>
</tr>
<tr>
<td>$x_d(7) = \bar{X}(11) = K_d$ [KN/cm]</td>
<td>4773</td>
<td>5901</td>
</tr>
</tbody>
</table>

**Performance: operational**

<table>
<thead>
<tr>
<th></th>
<th>Case 01</th>
<th>Case 02</th>
<th>Case 03</th>
<th>Case 04</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1(x_d)$</td>
<td>2.530</td>
<td>2.529</td>
<td>2.579</td>
<td>2.581</td>
</tr>
<tr>
<td>$\beta_2(x_d)$</td>
<td>2.400</td>
<td>2.263</td>
<td>2.399</td>
<td>2.459</td>
</tr>
<tr>
<td>$\beta_3(x_d)$</td>
<td>1.342</td>
<td>1.379</td>
<td>0.184</td>
<td>0.310</td>
</tr>
</tbody>
</table>

**Performance: life safety**

<table>
<thead>
<tr>
<th></th>
<th>Case 01</th>
<th>Case 02</th>
<th>Case 03</th>
<th>Case 04</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_4(x_d)$</td>
<td>3.969</td>
<td>4.017</td>
<td>4.175</td>
<td>4.127</td>
</tr>
<tr>
<td>$\beta_5(x_d)$</td>
<td>3.289</td>
<td>3.138</td>
<td>3.263</td>
<td>3.288</td>
</tr>
</tbody>
</table>

**Performance: collapse**

<table>
<thead>
<tr>
<th></th>
<th>Case 01</th>
<th>Case 02</th>
<th>Case 03</th>
<th>Case 04</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_6(x_d)$</td>
<td>4.063</td>
<td>4.600</td>
<td>9.152</td>
<td>10.14</td>
</tr>
<tr>
<td>$\beta_7(x_d)$</td>
<td>4.273</td>
<td>4.125</td>
<td>4.250</td>
<td>4.238</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Case 01</th>
<th>Case 02</th>
<th>Case 03</th>
<th>Case 04</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_{01}(x_d)$ initial cost, structure [ $ ]</td>
<td>1231800</td>
<td>1285700</td>
<td>1173600</td>
<td>1253600</td>
</tr>
<tr>
<td>$C_{02}(x_d)$ initial cost, controls [ $ ]</td>
<td>1917350</td>
<td>1944400</td>
<td>1738340</td>
<td>1645420</td>
</tr>
<tr>
<td>$C_0(x_d)$ total initial cost [ $ ]</td>
<td>11938150</td>
<td>12019100</td>
<td>11700940</td>
<td>11688020</td>
</tr>
<tr>
<td>$C_{d1}(x_d)$ repair costs, structure [ $ ]</td>
<td>0</td>
<td>14</td>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td>$C_{d2}(x_d)$ repair costs, controls [ $ ]</td>
<td>298850</td>
<td>164960</td>
<td>308690</td>
<td>333080</td>
</tr>
<tr>
<td>$C_{d3}(x_d)$ repair costs, non struc.[ $ ]</td>
<td>360</td>
<td>306</td>
<td>1031</td>
<td>1041</td>
</tr>
<tr>
<td>$C_{d4}(x_d)$ repair costs, contents [ $ ]</td>
<td>24550</td>
<td>24083</td>
<td>13287</td>
<td>13468</td>
</tr>
<tr>
<td>$C_5(x_d)$ social costs [ $ ]</td>
<td>149080</td>
<td>156330</td>
<td>40627</td>
<td>34905</td>
</tr>
<tr>
<td>$C(x_d)$ total cost [ $ ]</td>
<td>12410990</td>
<td>12364793</td>
<td>12064589</td>
<td>12070522</td>
</tr>
</tbody>
</table>

The difference in the total cost between the two cases is only 0.37%, showing that the cost function is relatively flat and that, therefore, the optimization algorithm converges to very similar minima when starting from different initial combinations, although there are
differences between the corresponding sets of design parameters.

Cases 03 and 04 in Table 3 correspond to situations in which the reliability constraints have not been imposed. The results for these cases are very similar, for all items in Table 3, either for costs or for reliability indices at the optimum solution, validating the robustness of the optimization algorithm.

A comparison between the case without reliability constraints and the case when such constraints are implemented shows that the latter results in higher structural cost $C_{01}$ (3.7%) and also a higher cost $C_{02}$ for the devices (14%). The reliability index $\beta_3(x_d)$, corresponding to the limit state of yielding in the dissipation devices, is now significantly lower when no reliability constraints are imposed. This implies a higher probability of yielding in the devices in case of an earthquake occurrence, leading to a need to replace the devices. However, all other reliability indices at the optimum solution do not differ substantially from the case when reliability constraints are imposed.

The repair costs for the dissipation devices, $C_{d2}$, is 28% greater (as an average of cases 03 and 04) than the costs corresponding to cases 01 and 02, when reliability constraints are imposed. Again, the passive dissipation devices protect the other structural components from damage. As a final result, the total cost results, on average, 2.7% higher when reliability constraints are introduced.

Should one opt for not imposing reliability constraints, the optimum would contain smaller dimensions and lower initial cost. However, there would be increased repair costs in case of an earthquake. If devices are not implemented, the repair cost of structural, non-structural elements and building contents would rapidly increase, along with social costs, leading to higher total costs.

5 CONCLUSIONS

An optimization process has been presented for the performance-based design of a reinforced concrete structure, including energy dissipation devices based on yielding. The objective of the optimization is to achieve a minimum total cost, including the initial cost of the structure, of the non-structural elements and contents, of the repair costs and social costs associated with the occurrence of earthquakes during the service life. The analysis considered a four-story office building. The optimization could be performed with or without minimum reliability constraints for each of the performance requirements.

When the reliability constraints are introduced, the optimum solution is controlled by the limit state of device yielding. The repair costs for damage to either structural or non-structural elements are very small or negligible, indicating that energy dissipation has been concentrated in the devices, which are replaceable, thus protecting the other building components.

When no reliability constraints are introduced, the initial cost of the devices is marginally smaller, but the repair and replacement cost after an earthquake increases rapidly. In comparison with the case when reliability constraints are introduced, the reliability indices are very similar, except the index associated with the limit state of device yielding.
REFERENCES


CONCEPTS OF DAMPERS FOR EARTHQUAKE PROTECTION OF EXISTING BUILDINGS AND FOR DISPLACEMENTS RESTRAINTS IN SEISMICALLY ISOLATED BUILDINGS

MIKAYEL G. MELKUMYAN

Armenian Association for Earthquake Engineering
1 Charents str., 0025, Yerevan, Armenia
mmelkumi@yahoo.com

Key words: Tuned Mass Damper, Isolated Upper Floor, Structural Concept, Earthquake Response Analysis, Dynamic Tests, First Application.

Abstract. Tuned Mass Damper (TMD) is one of the methods, known as a passive vibro-protecting device. The attempt to find the optimal parameters of TMD in the form of an additional flexible upper tenth floor (AFUF) in a 9-story frame building is presented. The efficiency of a single mass damper tuned to the first mode of building vibration is not very high. Therefore, three dampers tuned to the first three vibrations modes of the building are considered much more effective. However, solution of TMD in the form of AFUF contains some deficiencies, which are described in the paper. Therefore, the author has suggested providing flexibility to the damper using laminated rubber-steel bearings. In such case the AFUF will turn into an additional isolated upper floor (AIUF). Results of the earthquake response analyses and of dynamic testing of the existing 9-story apartment frame building before and after erection of AIUF are given.

Paper also presents a new concept of dynamic damper (DD) to restrict the displacements of seismically isolated buildings. Since the maximum displacement occurs at the level of isolators the proposed damper as a mass-spring subsystem attached either above or below the isolation interface. It is suggested to use the pavement around the building to create the damper in low-story structures. Other variant of damper is suggested for the large multistory buildings. In both cases the spring of the damper is represented by laminated rubber bearings, which work in one case under the tension and in another – under the compression forces. First real application of DD in construction of a seismically isolated residential house is described in the paper.

1 INTRODUCTION

There are more than 40 buildings in Armenia built, retrofitted or under construction employing the seismic (base and roof) isolation technologies, mostly using locally manufactured bearings with low or medium damping made of neoprene thus putting Armenia at a top rank in terms of the number of seismic isolated buildings per capita [1].

The effectiveness of an appended mass-spring system in reducing the dynamic response of a structure has been known for a long time. Numerous investigations and implementations of this idea for fixed-base buildings have been made [2,3,4]. In such buildings, the damper is
usually placed in an upper floor in order to experience a larger acceleration for efficiently mobilizing itself and absorbing the energy in the system.

In the case of a base isolated structure the damper must be attached immediately at the level of the isolation system. One option for reducing the displacement demand on the isolation system is to provide supplemental damping. This, however, may increase the in-structure accelerations [5]. A new type of damper to restrict the displacements of seismically isolated buildings called Dynamic Damper (DD) was developed and proposed by the author of this paper [6]. The name “dynamic damper” or “dynamic absorber” is applied when the auxiliary mass system has little damping [2]. It is suggested to use as the mass of the DD the perimeter pavement around the building which is separated from the superstructure and hung to it by means of laminated rubber bearings. The stiffness and the mass of the DD should be chosen such, that the period of vibration of the DD is equal to that of the isolated building. Such damper will allow to decrease the horizontal displacements and also to simplify the isolation system itself. The suggested structural system of the DD also increases the overturning resistance of the isolated building [6,7].

2 BACKGROUND AND LINEAR ANALYSES OF A BUILDING WITH AND WITHOUT TMD IN THE FORM OF AN ADDITIONAL UPPER FLOOR

Basically TMD is a single-degree-of-freedom appendage of the primary structure [8]. Dampers have been widely investigated in connection with seismic protection problems [8,9,10]. The natural frequency of TMD should be equal to the forced vibration frequency of the structure to be protected. Therefore, if the first vibration mode is the most significant one during earthquakes, then the natural frequency of the damper should be equal to the first mode frequency of structure vibration [11]. An additional upper floor for the buildings has been proposed as a vibration damper – TMD [12,13] and it could be erected on the existing buildings to increase their seismic resistance, without requiring the tenants to leave the building.

The attempt to find the optimal parameters of TMD in the form of an additional flexible upper tenth floor (AFUF) in 9-story frame buildings, using acceleration time histories of various earthquakes is presented below. The equations of the forced vibrations of a cantilever beam with masses concentrated at the floor levels are given by the formula [11]:

\[
m_k \ddot{y}_k + a_k (y_k - y_{k-1}) + \mu_k a_k (\dot{y}_k - \dot{y}_{k-1}) - a_{k+1} (y_{k+1} - y_k) - \mu_{k+1} a_{k+1} (\dot{y}_{k+1} - \dot{y}_k) = 0,
\]

where \( m_k, a_k, y_k \) are the mass, stiffness and displacement of the \( k \)th floor of the building, \( m = m_{10}, a = a_{10}, y = y_{10} \) are the mass, stiffness and displacement of the TMD-AFUF, \( \mu_k = \alpha / \omega \) is the coefficient of viscous damping of the \( k \)th floor, and \( \dot{y}_0(t) \) is the ground acceleration.

The values of floors’ stiffness and mass of the investigated building are as follows: \( a_1 = a_2 = \ldots = a_9 = 897000 \text{ kN/m}; \ m_1 = m_2 = \ldots = m_8 = 360 \frac{\text{kN} \cdot \text{s}^2}{m}; \ m_9 = 430 \frac{\text{kN} \cdot \text{s}^2}{m}. \) At \( \mu_k = 0 \) the periods of the three vibration modes of the building without TMD are equal to: \( T_1 = 0.778 \text{ s}, \ T_2 = 0.261 \text{ s}, \ T_3 = 0.159 \text{ s}. \) The building with TMD was analyzed using 12 acceleration time histories of strong earthquakes with the purpose to obtain the minimal values of the base shear.
forces, to determine corresponding optimal values of $\nu = m_2/m_1$ and $d = a_2/a_1$, and to compare the received results with those calculated for the building without TMD (Tab. 1).

**Table 1:** Optimal parameters of TMD and base shear forces of a 9-story building analyzed by 12 acceleration time histories with and without TMD

<table>
<thead>
<tr>
<th>Earthquakes</th>
<th>Optimal parameters determined for each time history</th>
<th>Base shear forces (kN) of the building</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\nu$</td>
<td>$d$</td>
</tr>
<tr>
<td>Ferndale, USA 7.10.1951, 44W</td>
<td>0.50</td>
<td>0.0150</td>
</tr>
<tr>
<td>Ferndale, USA 7.10.1951, 46E</td>
<td>1.25</td>
<td>0.0334</td>
</tr>
<tr>
<td>Ferndale, USA 21.12.1954, 44W</td>
<td>0.50</td>
<td>0.0100</td>
</tr>
<tr>
<td>Ferndale, USA 21.12.1954, 46E</td>
<td>1.00</td>
<td>0.0075</td>
</tr>
<tr>
<td>Ulcinj-2, Yugoslavia 15.04.1979, N-S</td>
<td>1.00</td>
<td>0.0075</td>
</tr>
<tr>
<td>Ulcinj-2, Yugoslavia 15.04.1979, N-E</td>
<td>1.25</td>
<td>0.0334</td>
</tr>
<tr>
<td>Herceg Novi, Yugoslavia 15.04.1979, N-S</td>
<td>1.00</td>
<td>0.0265</td>
</tr>
<tr>
<td>Herceg Novi, Yugoslavia 15.04.1979, N-E</td>
<td>1.25</td>
<td>0.0334</td>
</tr>
<tr>
<td>Ferndale, USA 3.10.1941, H60</td>
<td>0.75</td>
<td>0.0334</td>
</tr>
<tr>
<td>Hollister, USA 9.03.1949, H21</td>
<td>0.50</td>
<td>0.0100</td>
</tr>
<tr>
<td>Eureka, USA 21.12.1954, H10</td>
<td>0.50</td>
<td>0.0200</td>
</tr>
<tr>
<td>Taft, USA 12.01.1954, H70</td>
<td>0.50</td>
<td>0.0120</td>
</tr>
</tbody>
</table>

The results are showing that AFUF reduces the shear forces by about 35% in average. Seismic loads and shear forces, as well as displacements along the height of the building for both cases are shown in Figure 1. The mean values of the optimal parameters derived from Table 1 are the following: $\nu = 0.83$ and $d = 0.02$. Thus, the mass of TMD is equal to about 9% of the total mass of the building and its stiffness is about 50 times less than the stiffness of the building’s typical floor.

However, three dampers tuned to the first three vibrations modes of the building are considered much more effective and, therefore, a building structural solution with three TMDs has been proposed. When analyzing any building with TMDs, the number of vibration modes that should be taken into account is equal to the number of TMDs, with addition of at least the next three modes [10,11]. Namely, for the buildings with three dampers as it is schematically illustrated in Figure 2 at least six vibration modes should be encompassed in the analysis. The multi-version analyses of such structure allowed to conclude that in this case optimal stiffness and mass correlations of dampers could be found that enable significant reduction of shear forces and displacements (for about 2 times) compared to the building without TMDs. Reduction of lateral forces and displacements in the building with TMD takes place due to increase of vibration period of the whole system (building plus the TMD) and decrease of the first mode participation factors. However, a new type of second vibration mode appears and becomes prevailing, which results in the TMD oscillations in anti-phase relative to the building along the whole duration of the earthquake [12,13].
3 JUSTIFICATION OF TRANSITION FROM THE CONCEPT OF FLEXIBLE UPPER FLOOR TO THE CONCEPT OF ISOLATED UPPER FLOOR

The TMD in the form of AFUF considered above was implemented on the R/C 9-story frame building (Fig. 3). The AFUF represents a structure made of steel columns supporting a thick R/C slab. This building was tested using a powerful vibration machine installed on the slab of the 9th floor in two stages – before and after erection of AFUF [12]. Nevertheless, it became obvious that such a structural solution of AFUF contains some deficiencies from the practical point of view. In order to rigidly connect steel columns to the structural elements of the building, these columns should have sufficiently big cross-sections. But in this case the only way to provide the needed flexibility to the AFUF is to increase the height of steel columns (more than 4 m). However, this measure on one side reduces the resistance of AFUF against wind and on the other side raises its gravity center very high above the existing building. Therefore, during strong ground motions the flexible upper floor, though protecting the existing building, may itself suffer severe damages or even be destroyed causing damages to the building. Another deficiency is that no exterior and interior walls shall be constructed around and inside the space of the flexible floor as they will restrict its large horizontal displacement. Because of that and the possibility of partial or total destruction of AFUF during strong earthquakes it cannot be occupied and does not possess sufficient reliability.

All the above justifies the necessity to change the conceptual solution of this floor while keeping its idea. The author of this paper has suggested providing flexibility to the damper using laminated rubber bearings [10,13]. Obviously, in such case the known AFUF will turn into an additional isolated upper floor – AIUF (Fig. 4). Thus, the thin flexible columns are changed to seismic isolation LRBs and the slab, representing the mass of the flexible floor, is also changed to a whole upper floor connected to the existing building via laminated rubber bearings (LRBs).
It is important to note that in the proposed solution the R/C slab of AIUF is constructed right above the LRBs and comprises the largest portion of the damper-AIUF mass. Therefore, the gravity center of the damper in this new structural solution is very close to the existing building. Actually, AIUF itself above the isolation interface is a rigid structure, which being supported by LRBs undergoes practically no deformations during the earthquakes. Consequently, the suggested new concept of a TMD creation on top of the existing building allows not only to increase its seismic resistance and reliability of the whole system, but also to enlarge its useful space, which can be used for many different purposes.

4 NON-LINEAR SEISMIC RESPONSE ANALYSIS AND DYNAMIC TESTS OF THE 9-STORY FULL-SCALE EXISTING BUILDING BEFORE AND AFTER ERECTION OF AIUF

The method of AIUF was used in earthquake protection design and implementation for two existing R/C 9-story standard design buildings (Fig. 5). A special structure connecting the AIUF to the building was developed (Fig. 6). At the level of upper truss belts R/C slab is designed. The roof and the exterior walls of the AIUF were designed using light, “sandwich” type elements (Fig. 7). Free vibrations periods for a large number of this type of buildings were determined by the measurements of micro oscillations. The following results were obtained: first mode vibration period in transverse direction (along the R/C frames with weak beams and shear walls) \(T_{\text{trans}} = 0.48\) s in average, and in longitudinal direction (along the R/C frames with strong beams) \(T_{\text{long}} = 0.59\) s in average. Similar results for undamaged buildings are indicated in [14]. Design model of the building is presented in Figure 8.

Seismic response analysis was carried out for the building with and without AIUF, using degrading tri-linear model for columns and bilinear model for rubber bearings, as well as the Melkumyan model for shear walls [15], and using 7.12.1988, X direction Spitak Earthquake accelerogram scaled to 0.4g. The main results of non-linear seismic response analysis are given in Figure 9 and Table 2. The small scale of Figure 9 makes it hard to see the behavior of LRBs. Therefore, hysteresis loops for one LRB in a larger scale are presented in Figure 10. From the obtained results it can be seen that the R/C columns and shear walls of the building protected with TMD are mainly in the cracking stage, although yielding does occur in the shear walls of the lower four floors. Comparative analysis of the same building without TMD
shows that under the action of the same accelerogram the columns are in the yielding and shear walls are in the ultimate stages of deformation [10,11].

**Figure 5:** General views of the two existing R/C 9-story apartment buildings protected by AIUF

**Figure 6:** Steel rigid trusses constructed on the top of 9-story building and at the bottom of AIUF providing reliable connection of AIUF with the building by LRBs

**Figure 7:** The inner space of AIUF at the construction completion stage

**Figure 8:** Design model of 9-story building protected by AIUF

**Figure 9:** Restoring force - floor drift relationships for each floor of the building without and with AIUF

**Figure 10:** Force-displacement relationship for a rubber bearing of AIUF
Thus, the TMD provides sufficient earthquake protection to the building. The non-linear seismic response analysis proves that with TMD (AIUF) seismic loads experienced by the building could be reduced along the height of the building by about 2.5 times in average.

**Table 2**: The values of horizontal seismic shear forces and stages of deformation obtained by non-linear seismic response analysis of R/C 9-story apartment building with and without AIUF

<table>
<thead>
<tr>
<th>Story</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>AIUF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Seismic shear forces, kN</td>
<td>11601</td>
<td>11286</td>
<td>10589</td>
<td>9981</td>
<td>9548</td>
<td>9241</td>
<td>8803</td>
<td>7851</td>
<td>4723</td>
</tr>
<tr>
<td></td>
<td>Stages of deformation</td>
<td>In columns</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td>In shear walls</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
</tbody>
</table>

Building with AIUF

|       | Seismic shear forces, kN | 8332  | 8199  | 7154  | 6603  | 5130  | 4014  | 2927  | 1720  | 958  | 1315 |
|       | Stages of deformation | In columns | C | C | C | C | C | C | E | E | E |
|       | In shear walls | Y | Y | Y | Y | C | C | C | E | - | - |

E – elastic, C – cracking, Y – yielding, and U – ultimate stages of deformation

It was also decided to conduct dynamic tests of these buildings in two stages: first without AIUF, and then with it in resonance mode using unprecedented by its power vibration machine, which provided excitation of inertial horizontal loads allowing imitation of the design level seismic impact [13]. In the building test without AIUF the design intensity (VII by MSK-64 scale) was exceeded for about 6%. Along with that no damage was observed in the bearing structures. This means the building is capable to withstand reliably the intensity VII impact [10, 13].

Testing of the building with AIUF again was held in resonance regime, but in two vibration modes: AIUF and the building oscillate in the same phase (mode I/1), and AIUF oscillates in the anti-phase to the building (mode I/2). Comparison of the obtained shear forces at the ground floor level and displacements at the level of 9th floor slab have shown that thanks to the AIUF shear force and displacement are reduced by factors of 1.76 and 2.2, respectively. At the same time the drift of AIUF, or specifically the LRB displacement, exceeds the maximum drift of a story in the building by a factor of 4.3. However, this does not prevent using the AIUF space for various purposes, since its structures remain almost undeformed. That is why AIUF compares favorably with AFUF.

5 STRUCTURAL CONCEPTS OF DAMPERS FOR DISPLACEMENTS RESTRAINTS IN SEISMICALLY ISOLATED BUILDINGS

A 4-story base isolated apartment building constructed in Armenia in 1997-1998 is considered below as an example. For this base isolated building the location of a dynamic damper (DD) has been set at the level of the isolation system. With this purpose it was suggested to separate the blind area pavement around the building from the superstructure and then connect it to the superstructure via LRBs. Thus, the DD would consist of the mass (the mass of the pavement) and the spring (LRBs), and represent a single-degree-of-freedom system hung to the superstructure (Fig. 11).
However, the damper can be provided as a mass-spring subsystem attached either above or below the isolated floor of the building [10]. If the building represents a large multistory multifunctional system, which is different than the above mentioned apartment building, then of course the suggested dynamic damper could be modified. First of all, the damping factor for the isolation interface of the large building, as well as for the bearings of the damper will increase. In this case the dynamic damper will transform to a tuned mass damper (TMD). Second, its location in the building will require special consideration as the mass of such a TMD will be significantly bigger. For this case the suggested structural concept is shown in Figure 12. Obviously, that in contrast to the rubber bearings of the DD, which are working under the tension forces, the TMD’s rubber bearings will work under the compression forces.

One can imagine various functions for such a subsystem: an exercise room, a swimming pool, parking space, utilities room, as long as the mass remains relatively constant in time and large displacement can be accommodated. Proposed TMD scheme has the advantage of increasing the capacity of the base isolated building against overturning forces [7].

6 FIRST APPLICATION OF THE DYNAMIC DAMPER IN CONSTRUCTION OF SEISMICALLY ISOLATED RESIDENTIAL HOUSE

In order to identify the parameters of DD to be applied to a residential house it is assumed that a new mass is added to this system at the level of the isolation interface (Figs. 13, 14). Let us denote that $\omega_d$ is the frequency of vibrations of the damper’s added mass, $\beta = m_d/m$ – the ratio of the damper’s mass $m_d$ to the mass of the main isolated system $m$; $\omega_0$, $\omega_1$ and $\omega_2$ – the frequency of the main isolated system, and the first and second frequencies of the newly created system’s free vibrations, i.e. the main system plus the damper, and $n$ – is a damping factor. The problem is to choose the frequency of the damper $\omega_d$ so that at the given values of $\omega_0$, $n$ and $\beta$ the value of the displacement of the main isolated system after adding the DD would be minimal in comparison with its value before adding the DD for the whole duration of the earthquake. This is a very complicated task, which does not have a common solution for a system with an arbitrary initial frequency $\omega_0$, subject to an earthquake. For solving this problem it is necessary to choose a numerical value of $\beta$ for the given $\omega_0$, to present the frequency of damper $\omega_d$ in the form of $\omega_d = \nu \omega_0$ and to substitute it in the differential equation. This will allow determining $\omega_1$ and $\omega_2$ through the values $\nu$ and $\omega_0$. Then giving different
values to \( v \), it is possible to reach the minimum value of the displacement. The magnitude of factor \( v \), which corresponds to the minimum value of displacement, together with the value \( \beta \) will become a basis for designing the DD [16].

![Figure 13: Vertical elevation of the base isolated residential house and details of DD](image1)

![Figure 14: Three-dimensional view of the seismic isolation system and the DD](image2)

It was suggested to use the space under the perimeter pavement around the building for creation of DD connecting it to the pavement by LRBs installed with certain spacing. The mass of DD in its turn is designed as a continuous R/C beam around the whole perimeter of the building (Fig. 15). This mass in case of a low-story building should be around 3-5% of the mass of the superstructure. As it is mentioned above, the isolated building with DD will have two main modes of vibrations: when the DD oscillates in the same phase with the building and when it oscillates in anti-phase to the building. It is this second mode that becomes prevailing and due to this phenomenon horizontal displacements and forces are reduced.

![Figure 15: Pavement around the building before placing the reinforcement and casting the concrete (a) and the final view of DD as a continuous beam hung to the pavement by means of laminated rubber bearings (b)](image3)

7 NON-LINEAR EARTHQUAKE RESPONSE ANALYSIS OF THE RESIDENTIAL HOUSE WITH AND WITHOUT DYNAMIC DAMPER

The building was analyzed by SAP2000 non-linear program by a three-dimensional design model (Fig. 16) using the above mentioned acceleration time history recorded at Ashotsk station. Experimentally obtained characteristics of seismic isolation rubber bearings were as follows: total initial stiffness of sixteen rubber bearings – 1365.8 kN/cm, strain hardening (effective) stiffness – 129.6 kN/cm, yield strength – 273.2 kN, yield displacement – 0.2 cm.
The analysis was carried out for different values of the DD weight (or $\beta$ and $\nu$) in order to identify the most effective parameters leading to maximum reduction of the isolation system displacement. The range of $\beta$ (the ratio of the DD weight to the weight of the building) varied from 0.42% to 7.01%. Some results of analysis of the building with and without DD are given in Table 3 and in Figure 17.

![Figure 16: 3D design model of the base isolated residential house](image1)

![Figure 17: Horizontal displacements-factor $\beta$ relationships for the structure with and without DD](image2)

**Table 3:** Results of the non-linear earthquake response analysis of the residential house with and without DD

<table>
<thead>
<tr>
<th>$W_{DD}$, weight of DD, kN</th>
<th>$W_{DD}/W_S$, %</th>
<th>$Y_S$, displacement of structure, mm</th>
<th>$Y_{DD}$, displacement of DD, mm</th>
<th>$T_1$, period of the first mode, sec</th>
<th>$T_2$, period of the second mode, sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without DD, weight of structure $W_S = 13900$ kN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>127.14</td>
<td>-</td>
<td>2.07</td>
<td>-</td>
</tr>
<tr>
<td>974.98</td>
<td>7.01</td>
<td>70.37</td>
<td>115.50</td>
<td>2.80</td>
<td>1.97</td>
</tr>
<tr>
<td>882.18</td>
<td>6.35</td>
<td>67.21</td>
<td>112.51</td>
<td>2.72</td>
<td>1.96</td>
</tr>
<tr>
<td>710.50</td>
<td>5.11</td>
<td>62.93</td>
<td>107.58</td>
<td>2.57</td>
<td>1.93</td>
</tr>
<tr>
<td>612.63</td>
<td>4.41</td>
<td>62.87</td>
<td>103.69</td>
<td>2.48</td>
<td>1.90</td>
</tr>
<tr>
<td>522.00</td>
<td>3.76</td>
<td>62.82</td>
<td>99.98</td>
<td>2.41</td>
<td>1.87</td>
</tr>
<tr>
<td>435.62</td>
<td>3.13</td>
<td>65.85</td>
<td>107.79</td>
<td>2.35</td>
<td>1.83</td>
</tr>
<tr>
<td>362.50</td>
<td>2.61</td>
<td>72.40</td>
<td>128.37</td>
<td>2.30</td>
<td>1.78</td>
</tr>
<tr>
<td>293.62</td>
<td>2.11</td>
<td>80.00</td>
<td>149.18</td>
<td>2.26</td>
<td>1.73</td>
</tr>
<tr>
<td>232.00</td>
<td>1.67</td>
<td>90.16</td>
<td>166.21</td>
<td>2.23</td>
<td>1.67</td>
</tr>
<tr>
<td>177.62</td>
<td>1.28</td>
<td>98.62</td>
<td>177.62</td>
<td>2.21</td>
<td>1.61</td>
</tr>
<tr>
<td>130.50</td>
<td>0.94</td>
<td>105.54</td>
<td>183.82</td>
<td>2.19</td>
<td>1.56</td>
</tr>
<tr>
<td>90.62</td>
<td>0.65</td>
<td>110.80</td>
<td>186.29</td>
<td>2.18</td>
<td>1.51</td>
</tr>
<tr>
<td>58.00</td>
<td>0.42</td>
<td>114.56</td>
<td>186.56</td>
<td>2.17</td>
<td>1.46</td>
</tr>
</tbody>
</table>
The obtained results show that the suggested structural concept of DD is very effective and leads to significant reduction of the seismically isolated structure’s horizontal displacements. Maximum reduction occurs at the values of factor $\beta$ (WDD/WS) varying from 3.76% to 5.11%. For this range of factor $\beta$, the factor $\nu$ ($\omega_{DD}/\omega_0$) in average is equal to 1.4. DD with these parameters reduces the horizontal displacements and the shear forces at the level of isolation system by 2 times in average. Table 3 also suggests that the building’s period of the first mode of vibrations without DD is between the periods of the new first and second modes of vibrations of the building with DD and differs from them by +0.42 sec and -0.17 sec.

8 CONCLUSIONS

- Tuned mass dampers as additional upper floors in the form of AFUF or AIUF are suggested and presented. The efficiency of a single mass damper tuned to the first mode of building vibration is not very high. Three dampers tuned to the first three vibrations modes of the building are considered much more effective as in this case optimal stiffness and mass correlations of dampers could be found that enable significant reduction of shear forces and displacements (for about 2 times) compared to the building without TMD.

- Deficiencies of AFUF are described and, thus, providing flexibility to the damper using LRBs is suggested. Transition from the concept of AFUF to the concept of AIUF is justified. The non-linear seismic response analysis proves that with AIUF, acting as a TMD, seismic loads (the strain-stressed state level) experienced by the building could be reduced along the height of the building by about 2.5 times in average.

- Dynamic testing of the existing 9-story building before and after erection of AIUF allows to conclude that the proposed AIUF method leads to upgrading earthquake resistance of buildings and that AIUF brings to reduction of shear force at the ground floor level by a factor of 1.76 and at the same time the displacement at the 9th floor slab level decreases 2.2 times.

- The proposed new concept of a damper as a mass-spring subsystem attached either above or below the isolation interface is an effective means to restrict the displacements of seismically isolated buildings. The spring of the damper is represented by LRBs. Non-linear earthquake response analyses of the base isolated buildings with and without DD show that due to application of the suggested structural solutions of DD significant (about 2 times) reduction of the horizontal displacements and shear forces at the level of isolation system takes place. First real application of DD in construction of a seismically isolated residential house is described.

REFERENCES


ATOMISTIC STUDY OF THE MECHANICAL PROPERTIES OF A SINTERED BULK METALLIC GLASS (NANOGLASS)

FRANCO ARDIANI¹, ANDRÉS A. MANELLI¹, CARLOS J. RUESTES², CLAUDIO A. CAREGlio¹,³ AND EDUARDO M. BRINGA²,⁴

¹Facultad de Ingeniería, Universidad Nacional de Cuyo
Centro Universitario, 5500 Mendoza, Argentina
francoamg@gmail.com, andresmanelli@gmail.com, ccareglio@uncu.edu.ar

²ITIC, Universidad Nacional de Cuyo
Centro Universitario, 5500 Mendoza, Argentina

³FCEN, Universidad Nacional de Cuyo
Centro Universitario, 5500 Mendoza, Argentina
cjruestes@hotmail.com, diego.tramontina@gmail.com, ebringa@yahoo.com

⁴CONICET
5500 Mendoza, Argentina

Key words: Metallic Glasses, Large Deformations, Porosity, Atomistic Analysis

Abstract. Metallic glasses with porosity have been in the spotlight in recent years and much research has been done on them, in an effort to improve the understanding of the mechanics of deformation. The behavior of the material in the elastoplastic regime may be controlled by the introduction of pores, thus the interest on simulating them.

Large deformation is normally a consequence of shear transformations zones (STZ) collapsing into shear bands (SB) which may lead to catastrophic failure. It is widely known that in crystalline metals the addition of nanopores slows down dislocation motion and changes the resulting plastic deformation. Similarly, pores in metallic glasses limit the propagation of shear bands and allow a more homogeneous deformation.

In previous work we determined constitutive parameters of the Cu₄₆ Zr₅₄ metallic glass as a function of temperature. We now present results for a metallic glass with the same composition, but fabricated by sintering of BMG nanoparticles, which results in samples with porosity, similar to nanoglass samples in some experiments and simulations. Atomistic simulations are carried out using Molecular Dynamics (MD), and analysis includes atomic Voronoi polyhedra, and atomic stress and strain. We analyse the dependence of deformation on solid volume fraction, and the way deformation is distributed throughout the sample, as a function of initial porosity.
1 INTRODUCTION

A Metallic Glass (MG), also called “amorphous metal”, is a metallic alloy which exhibits an amorphous structure, contrary to the most usual crystalline structure of metals. This may be accomplished by several techniques, most of which include high quenching rates, low volumes and composition control [1]. As a result, these materials have some advantages over crystalline metals, such as improved elasticity combined with high resistance, strength and moldability [2].

There are two main approaches to computationally simulate MGs under plastic strain: focusing on nanoscale behavior [3, 4] or using continuum mechanics [5]. For the first approach, Molecular Dynamics (MD) simulations are often used [6]. MD solves problems with many bodies by applying an atom to atom potential, thus being useful to study nanoscale properties, such as strain, stress, temperature, etc.

MD simulations are also useful to identify plastic processes in BMGs (Bulk Metallic Glasses). Plasticity in BMGs starts with the formation of Shear Transformation Zones (STZ), which nucleate into shear bands [3, 7] with increasing strain. Shear bands (SB) may lead to brittle failure of the material due to heterogeneous deformation, thus the importance of preventing their propagation. In crystalline metals a similar thing is done with dislocations.

A more homogeneous deformation may be achieved by adding nanoinclusions to the material. There has been great interest in this subject lately, and many options have been explored [8, 9, 10, 11]. Following previous work [12], we now fabricate Cu$_{46}$Zr$_{54}$ BMG samples with porosity (similar to nanoglass samples in other experiments and simulations [13, 14]) by sintering of nanoparticles. MD simulations will allow us to analyse atomic Voronoi polyhedra, atomic stress and strain, and the way initial porosity affects the deformation of the sample.

2 SIMULATION DETAILS AND POROUS SAMPLE PREPARATION

For this paper, MD simulations were carried out using the LAMMPS software [15], which is free and open source, has an excellent manual, and is computationally efficient in the simulation of systems with large numbers of atoms. In addition Voronoi analysis and images of the sample were made with Ovito software [16], and other figures were plotted with Gnuplot, both of these softwares also being free and open source.

For the preparation of the porous sample, we started from the sample used in previous work [12] and described by Arman [17]. It is a Cu$_{46}$Zr$_{54}$ prismatic sample with a total of about 160000 atoms and obtained with a cooling rate of $10^{12}$ K/s. The experimental glass transition temperature ($T_g$) of this metallic glass is 696 K, and the experimental shear modulus (G) is 30 GPa [18]. To describe the interactions between atoms, an embedded atom method (EAM) potential [19] is adopted. We use 3D periodic boundary conditions, suitable for high strain rates [20], so as to simulate a BMG and avoid stress concentration at the boundaries.
Figure 1: Images of the sample (a) prior to sintering process and (b) after sintering process (13% porosity).

We took the original sample and replicated it along one direction so as to obtain a roughly cubic sample, with a side of about 15 nm. Afterwards, we randomly selected points within this sample which would be the center of spheres (2.5 nm radius), and remove all atoms outside those spheres in order to simulate sintering of spherical glass
nanoparticles. Figure 1-(a) shows the sample prior to the sintering process. This sample has 77888 atoms.

The procedure for the simulated sintering was to relax the sample at a constant high-temperature of 650K, just below the glass temperature, at constant volume during few ps, and then apply up to 10 ps of compressive pressure (400 bar). We then repeated these two steps until the desired porosities were obtained. Then, we performed further relaxation with the following procedure: cooling from the sintering temperature to zero temperature at a rate of $6.5 \cdot 10^{14} K/s$, applying a barostat to reach zero pressure, heating at the same rate than the cooling rate to arrive to the desired simulation temperature (300K) and, finally, applying a barostat during 5 ps to reduce pressure to zero while maintaining constant temperature. Figure 1-(b) shows one of the samples that result from the mentioned sintering procedure.

Samples with different initial porosities (3.3%, 5.8% and 13.1%) were prepared. These stable samples were then used to perform uniaxial compressive and tensional loading. All atomic coordinates were scaled every step, according to the desired strain rate, which in this case was $10^9/s$, appropriate for shock compression experiments.

3 RESULTS

Below we present results for purely uniaxial strain, which is adequate for the comparison with results of experiments at very high strain rates, where lateral strains can be neglected during loading.

Figure 2-(a) plots Pressure along the Z-axis (loading axis) versus strain for compressive strain. The presence of porosity helps the initiation of plasticity, due to pores acting as shear concentrators, thus facilitating the appearance of STZs and shear bands. This early plasticity starts to close out the pores, producing a curve where the strain increases while maintaining low pressure. When the pores close, the pressure increases in a more accelerated manner, as shown by the portion of curve after the dashed lines. Moreover, we can as well observe that the behavior of the porous curves after the dashed lines matches the behavior of the non-porous curve, implying there is no more porosity. Based on the figure, we can conclude that, with increasing porosity, less pressure is needed to close the pores (height of the dashed lines), but this happens at a higher strain.

Under tensile strain, the sample behaves quite differently. We have found, from the analysis of snapshots of the sample (as those seen in Figure 5), that pores do not close under tension as they did under compression. What is more, they seem to grow at a roughly constant rate with increasing strain. The use of periodic boundary conditions precludes the closing of the voids even at 20% uniaxial strain, given that there is no lateral strain. Figure 2-(b) shows what seems to be plastic flow: particularly for 13% porosity, but seemingly for the other porosities, after a certain point, strain increases while pressure stays constant or even decreases.

Figure 3 shows Voronoi polyhedra-strain curves. The Voronoi tesselation analysis is a technique for characterizing short range order in amorphous metallic glasses, where each
atom is the center of a Voronoi polyhedron, formed of its nearest neighbors. In [17], type 3 atoms are identified as plasticity indicators, thus their importance.

Figure 3-(a) shows Voronoi polyhedra for compressive strain. This figure shows a fall in the number of Type 3 atoms after a constant stage. This has been thought to be an indicator of the onset of plasticity [17]. However, our graphic shows a counter-intuitive result, as plasticity starts earlier in the samples with less porosity according to this Voronoi
analysis. This could be considered as an indicator that there are other processes involved in early stages of deformation that have an influence in the results.

Figures 4-(b) and (c) show the evolution of shear strain in the sample, for compressive strain. We can easily observe that pores act as stress concentrators, but they also represent an obstacle for shear band propagation [9]. Shear bands form diagonally from pore to pore, and the atomic strain continues to accumulate along this directions for the rest of

Figure 3: Type 3 Voronoi polyhedra-strain curves. (a) Compression (b) Tension.
Figure 4: Shear strain coloring of the 13% porosity sample’s slice. The shear strain coloring was made using ovito, blue is 0.1 or less and red is 0.3 or more shear strain. (a) 0% strain (b) 5% compressive strain (c) 12% compressive strain.

the loading, as shown by the bulk glass at 12% strain. Hardening appears moments before the total closure of the pores as has been pointed out by [21] and can be appreciated in Figure 2-(a).

Figure 3-(b) shows Voronoi polyhedra for tensile strain. In this figure, type 3 atoms show practically no variation for the porous samples, which would imply that no STZs are
Figure 5: Shear strain coloring of the 13% porosity sample’s slice. The shear strain coloring was made using ovito, blue is 0.1 or less and red is 0.3 or more shear strain. (a) 6% tensile strain (b) 20% tensile strain.

being formed. For the non-porous sample, the count of type 3 atoms becomes somewhat constant after the void has been nucleated. This led us to think that, given the conditions for the samples, the movement of the atoms around the pores is facilitated, and prevents the formation of STZs other than the ones around the pores. To support this idea, in Figures 5-(b) and (c) we show the sample colored for shear strain. It is evident that shear strain is mostly concentrated around the pores. It should be mentioned that the relative position between atoms far from the pores remains almost the same.

4 CONCLUSIONS

Molecular Dynamics (MD) simulations were carried out on a porous Cu_{46}Zr_{54} metallic glass sample, applying compressive and tensile loading. The results under strain were comparable to those found in the literature [21] for compresion of porous monocrystalline Cu samples. This could be considered as validation of the sintering process used to prepare the nanoglass samples.

Under compression, pores facilitate the plastic process acting as stress concentrators, but they delay as well the formation of Shear Transformation Zones (STZs) and their possible union into a full shear band (SB), for the material far from the pores. Results also show hardening of the sample following the total closure of the pores, very much similar to the non-porous case.

Under tension, and using purely uniaxial strain, pores do not close and they concentrate plastic flow around them, also impeding the formation of STZ and SBs.
Future study will also include a deeper Voronoi analysis and simulations of larger samples, with different void topologies.

5 ACKNOWLEDGEMENTS

We all thank support from SeCTyP-UNCuyo grant “B008 Study of possible improvements of the mechanical properties of metallic glasses”. EMB and CJR also thank support from PICT-PRH 0092.

REFERENCES


MECHANICAL PROPERTIES OF A CU$_{46}$ZR$_{54}$ BULK METALLIC GLASS WITH EMBEDDED CRYSTALLINE NANO PARTICLES

ANDRES A. MANELLI$^1$, FRANCO ARDIANI$^1$, CLAUDIO A. CAREGlio$^{1,2}$ AND EDUARDO M. BRINGA$^{3,4}$

$^1$Facultad de Ingeniería, Universidad Nacional de Cuyo
Centro Universitario, 5500 Mendoza, Argentina
andresmanelli@gmail.com, francoamg@gmail.com, ccareglio@uncu.edu.ar

$^2$ITIC, Universidad Nacional de Cuyo Centro Universitario, 5500 Mendoza, Argentina

cjruestes@hotmail.com, diego.tramontina@gmail.com, ebringa@yahoo.com

$^3$FCEN, Universidad Nacional de Cuyo
Centro Universitario, 5500 Mendoza, Argentina

$^4$CONICET
5500 Mendoza, Argentina

Key words: Computational Mechanics, Metallic Glass, Inclusions, Diffusivity

Abstract. Plasticity in bulk metallic glasses (BMGs), is normally dominated initially by shear transformations zones (STZ), which expand to form shear bands (SB) through the material. In order to control and thus improve the dynamics of plasticity, composition of metallic glasses has been modified in different ways.

Particularly, the inclusion of crystalline nanoparticles provides obstacles to SB propagation and growth, with SB often nucleating at the interface between the BMG and the nanoparticle. This results in a reduced and more homogeneous deformation in the plastic regime. Nevertheless, to ensure lasting effects, inclusions should be stable in time, i.e. not diffuse into the surrounding amorphous material loosing the sharp transition from crystal to amorphous.

In previous work we determined constitutive parameters of the Cu$_{46}$Zr$_{54}$ metallic glass as a function of temperature, using atomistic Molecular Dynamics (MD) simulations. We will now present results for spherical face-centered cubic (FCC) Cu inclusions. Although we do not focus on the size effects of inclusions like other studies, we analyze the stability of the nanoparticles at different temperatures. During mechanical deformation under uniaxial strain of a BMG sample with inclusions, we analyze Voronoi polyhedra, and shear stress and shear strain localization to study the role of the inclusion in the mechanical properties of this composite material.
1 INTRODUCTION

Bulk Metallic Glasses (BMG) are known for their outstanding mechanical properties such as elasticity, strength and hardness, and because of this they are matter of deep research. However, the tendency to massively collapse initially formed shear transformation zones (STZs) into a single shear band (SB) and thus failing catastrophically, restricts their plasticity and their applications [1, 2].

It is of much interest the modification of their composition so as to enhance these restraints [3, 4, 5]. In particular, the inclusion of crystalline nanoparticles promote the nucleation of STZs and act as an impediment to the propagation of SBs [3]. As a result, a reduced and more homogeneous deformation in the plastic regime is obtained.

After analyzing constitutive parameters of the Cu$_{46}$Zr$_{54}$ metallic glass used as matrix in this work [6], we now focus on the stability at different temperatures of spherical face-centered cubic (FCC) Cu inclusions using Molecular Dynamics (MD) simulations. Voronoi polyheadra, shear stress and shear strain are analyzed in the deformation process under uniaxial strain of the BMG in order to get a better understanding of the role of the inclusion in the mechanical properties of this composite material.

2 SIMULATION DETAILS

We use the LAMMPS software package [7] for running all the simulations, which is free and open source. The original sample is a Cu$_{46}$Zr$_{54}$ metallic glass with 160k atoms, obtained with a quenching rate of $10^{12}$ K/s with an experimental glass transition temperature ($T_g$) of 696 K, and it has already been described in [8]. An embedded atom method (EAM) potential is adopted [9], used previously in other works on BMGs [1, 8, 10, 11, 12, 13]. Periodic boundary conditions are used in all directions, to mimic high strain rate loading conditions. We analyze FCC Cu sphere-shaped inclusions. A spherical region of 2 nm radius was deleted from the sample in a central position which was then filled with the corresponding lattice, i.e. an FCC lattice with lattice constant of 0.3615 nm. After creating Cu atoms in the spherical region, the configuration was minimized, then equilibrated at zero pressure for few ps, then heated (or cooled) to reach the desired final temperature ($T_f$) during 4 ps and it was finally annealed at $T_f$ for 1 ns.

The present work centers on the stability of these nanoparticles at different temperatures, although we show some cases of uniaxial tension and compression loading. Homogeneous strain rate of $10^9$/s is applied. Diffusivity in all cases is computed fitting mean squared displacements (MSD), $\langle r^2 \rangle$, output from LAMMPS.
3 RESULTS

3.1 Inclusion Stability

In this section we analyze the stability of the nanoparticles at different temperatures, up to temperatures close to the glass transition.

MSD data for Cu atoms from the inclusion alone are obtained from simulations at different temperatures. As we can see in Figure 1 (a), after an initial transient behavior, the MSD becomes nearly constant, leading to a diffusivity close to zero, as expected for a stable solid inclusion.

This would indicate that the inclusions are indeed stable at normal operational conditions for BMGs. However, MD simulations are typically cover relatively short time spans of only few ns, and we move to higher temperatures for an improved assessment of nanoparticle stability.

We then use data for $T \geq 500$ K (Figure 1 (b)), and obtain diffusivities shown in Table 1, using Einstein’s equation $\langle r^2 \rangle = 6Dt$, for times longer than 0.6 ns, when there is a steady slope.

These results represent an average for all nanoparticle atoms, but there could be large differences between atoms in the core and the surface of the particle. Therefore, we calculate the MSD for Cu atoms within a spherical shell of thickness 0.8 nm, and inner radius of 1.2 nm. It can be seen in Figure 2 that, after a transient of around 0.6 ns, the slopes of the MSD are roughly the same for the shell and the entire nanoparticle, and we can use the diffusivities in Table 1.

Diffusivities for $T \geq 500$ K are fitted to match equation 1, where $k_B$ is the Boltzmann’s constant and $\Delta E$ is the activation energy for diffusion, and $D_0$ sets the diffusivity scale. The resulting regression results appear in Table 2 and are displayed in Figure 3.

Table 1: Diffusivity fitting results

<table>
<thead>
<tr>
<th>$T$ [K]</th>
<th>$D$ [nm$^2$/ps]</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>$8.490 \cdot 10^{-7}$</td>
<td>0.8306</td>
</tr>
<tr>
<td>600</td>
<td>$1.508 \cdot 10^{-6}$</td>
<td>0.9253</td>
</tr>
<tr>
<td>700</td>
<td>$4.699 \cdot 10^{-6}$</td>
<td>0.9357</td>
</tr>
<tr>
<td>800</td>
<td>$4.149 \cdot 10^{-6}$</td>
<td>0.9935</td>
</tr>
</tbody>
</table>

\[ D = D_0 \cdot e^{-\Delta E/k_B T} \]  

(1)

As pointed out in [3], the crystalline core of inclusions shrinks because the external atoms tend to became amorphous. This effect increases with temperatures, as expected, when atomic diffusion is larger.
Figure 1: MSD for FCC Cu inclusion at different temperatures

(a) Lower temperatures

(b) Higher temperatures

Figure 1: MSD for FCC Cu inclusion at different temperatures
Figure 2: MSD of spherical shell and original for the whole sphere (lighter color)

Figure 3: Diffusivity in function of temperature
Table 2: Diffusivity with respect to T fitting results

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Activation energy [eV]</td>
<td>−0.4182</td>
</tr>
<tr>
<td>$D_0 \left[ \text{nm}^2 \text{ps}^{-1} \right]$</td>
<td>$8.771 \times 10^{-3}$</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.8399</td>
</tr>
</tbody>
</table>

3.2 Uniaxial Loading

In this section, we present results on uniaxial loading of the BMG containing a nanoparticle, as described before.

Figures 4 and 5 show the sample under uniaxial stress (tension and compression) for $T = 10$ K and $400$ K. We consider these two temperatures as representative of a low temperature regime and a high temperature regime, but for temperatures for which diffusivity is small and the nanoparticle is stable.

The shape of the loading curves follows the expected behavior observed for the material without an inclusion [6]. The small inclusion does not affect the elastic regime, and at high temperature, the softening of the BMG is not affected either, neither in tension nor in compression. The maximum shear is slightly decreased by the nanoparticle at low temperatures. The flow stress at large strains is not modified by the inclusion under compression. On the other hand, the nucleation of a void under tension, indicated by the sudden drop in the stress, is slightly delayed by the inclusion, allowing about 1% additional strain of the sample. This might be due to some relaxation and dissipation taking place in the boundary between the nanoparticle and the BMG, but further studies are needed to clarify this.

Under tension (Figure 4), even if atomic shear strain seems to be concentrating at the boundaries of the inclusion located at the central zone of the sample as shown in Figure 6, homogeneous plastic strain is present, with nucleation of many STZs, probably due to the high strain rate and high quenching rate of the sample, leading to void nucleation in a zone different from the boundary between the BMG and the nanoparticle. Under compression at low temperature, there are also STZs, as shown in Figure 7.

As some particular Voronoi polyhedra are considered to be more shear resistant structures, particularly icosahedra clusters [14], we show the evolution of icosahedral fractions for the sample and compare it with the original BMG with no inclusion at 10 K. We can see in Figure 8 that fractions follow the same behavior as in the sample without a nanoparticle. Under tension, void nucleation leads to fluctuations. It is worth noting that these fluctuations correspond to a higher strain like we already mentioned for the stress-strain curve.
Figure 4: vonMises stress vs strain for the BMG under uniaxial tension with no inclusion (dotted line) and one inclusion (solid line).

Figure 5: vonMises stress vs strain for the BMG under uniaxial compression with no inclusion (dotted line) and one inclusion (solid line).
Figure 6: Snapshots of sample under tension at selected strains at 10 K (inclusion pointed out in first snapshot). Color represents atomic shear strain.

Figure 7: Snapshots of sample under compression at selected strains at 10 K (inclusion pointed out in first snapshot). Color represents atomic shear strain.
Figure 8: Icosahedral fraction under tension and compression at 10 K. Dotted line represents original sample.

4 SUMMARY AND CONCLUSIONS

We study a Bulk Metallic Glass with a crystalline nanoparticle as inclusion. We consider a CuZr glass, and a pure Cu nanoparticle with a radius of 2 nm. This implies a volume fraction that varies from 1.15% at 10 K to 1.12% at 800 K as a result of the increasing initial volume of the sample with temperature. A similar scenario was recently explored by Albe et al. [3]. Here we explore temperature effects, and initially we study the thermal stability below 400 K, indicating that the nanoparticle is fairly stable at those temperatures. At higher temperatures, diffusivity within few ns leads to the loss of a sharp interface between the nanoparticle and the glass.

In our simulations 3D-periodic boundary conditions were applied, and the absence of free surfaces leaves only the copper precipitate as likely stress concentrator to promote nucleation of STZs, as seeds for shear bands, at the interface between the nanoparticle and the glass. However, this was not the case. Stress-strain curves are fairly similar to the case without the nanoparticle, with the exception of a delay in void nucleation under tension for the sample with a nanoparticle.

Voronoi analysis did not show significant differences between the samples with and without the nanoparticle. Further voronoi analysis is required for different loading temperatures and loading conditions. Future studies will also repeat these experiments with
CuZr inclusions with a crystalline B2 structure, as found in some experiments [15, 16].

5 ACKNOWLEDGMENTS

We all thank support from SeCTyP-UNCuyo grant “B008 Study of possible improvements of the mechanical properties of metallic glasses”. Eduardo M. Bringa also thanks support from PICT-PRH 0092.

REFERENCES


NUMERICAL SIMULATION OF ELASTIC SPRINGBACK IN PROBLEMS INCLUDING LARGE ELASTIC-PLASTIC STRAINS

WALTER B. CASTELLÓ¹ and FERNANDO G. FLORES²

¹Departamento de Estructuras, Universidad Nacional de Córdoba
Casilla de Correo 916, 5000 Córdoba, Argentina
wcastello@efn.uncor.edu

²Departamento de Estructuras, Universidad Nacional de Córdoba and CONICET
Casilla de Correo 916, 5000 Córdoba, Argentina
fflores@efn.uncor.edu

Key words: finite elements, large deformations, springback

Abstract. Deep drawing of sheets, is one of the most used manufacturing processes for metal forming. In these processes, the proper design of the tools geometry allows the fabrication of very complex parts with a high level of precision. The formed blank may exhibit defects such as wrinkles, thinning and metal damage, also they may present important differences with the target geometry due to elastic recovery or springback. In this paper, the elastic-plastic two-surface model of Yoshida-Uemori is used to study the influence of springback in deep drawing problems. The two-surface constitutive model allows taking into account the variation of the elastic modulus of the material as a result of the undergoing plastic strains. An elastic-plastic model with linear isotropic hardening including changes in the elastic modulus has also been considered in order to compare the results with the previous two-surface model. Experimental results show that the use of a plastic strain dependent elastic modulus may allow to adequately capture the geometry of the piece after springback. The results obtained in this study agree with those obtained experimentally in and numerically by the authors. It has also been found that when the kinematic-hardening is not high the model with only isotropic hardening adequately captures the springback phenomenon.

1 INTRODUCTION

Presently the geometric precision is a primary requirement in the metallurgical industry to decrease the amount of subsequent operations to forming of metal parts. In the sheet stamping and forming processes one of the main aspects to consider is springback
phenomenon, which must be counterbalanced to obtain a product with adequate precision. In the works [1, 2, 3] an overview of developments in the simulation of springback is presented, particularly in [3] it is observed that studies in this area have increased in the last decade in the following items: a) constitutive equations, b) change in Young’s modulus, c) through the thickness stress integration. The first two items are related to the constitutive model used, while the third depends on the numerical approximation chosen for analysis.

Historically the numerical approximation schemes used were based on shell theories. However, additional developments have been considered to assess springback such as: the improved bending approach of [4] and those mentioned therein; or more robust schemes to compute the stresses across the thickness [5] as they are critical in predicting the springback. Due to the importance of the computation of stresses across the thickness, it is more appropriate to use approximations based on solid or solid-shell elements. These approaches provide additional advantages in the study of drawing and stamping problems as set forth in [7]. However they have disadvantages, such as: 1) a stiffer bending behavior, 2) a larger number of Gauss points, and 3) the time increment in explicit integration schemes. Regarding points 1) and 2), use of elements with reduced integration and/or assumed strains as can be seen in [6, 7, 8, 9, 10] have led to important improvements. In [6] an enhanced assumed strains (EAS) solid element with a single point of integration is used. In [8, 10] a one point integration element plus stabilization of hourglass modes using assumed natural strain (ANS) technique based on a B-bar type method is used. In [7] an eight-node solid element which employs ANS for transverse strains is considered. While in [9] an ANS and EAS six-node prismatic element where neighboring elements are used for strain evaluation is developed. Finally, according to [7] and [10] an efficient way to alleviate the problem of small time step in explicit integration schemes is the selective mass scaling.

Initially springback was estimated assuming a completely elastic recovery, however it has been shown that inelastic strains produce an additional recovery [11]. The accuracy in the evaluation of these inelastic strains is linked to the constitutive material model used. In [12] the influence of different constitutive models in the springback of sheets have been studied, with particular emphasis on the effect of the type or hardening of the model. The constitutive models for metals have rapidly evolved in the last three decades, an extensive comparison between different theoretical frameworks can be found in [13]. Among the most used constitutive models to predict springback are multiple yield surfaces models, including Yoshida-Uemori model [14, 15] that represents an interesting option due to the relatively small number of parameters used to define it.

In this work the finite element code Simpact with explicit integration of governing equations is used. The eight-node solid element proposed in [7] is considered. This element has an ANS for transverse strains and only four integration points that makes it efficient and robust to deal with sheet springback problems. Besides the two-surface model of Yoshida-Uemori has been implemented, which allows taking into account both
isotropic and kinematic hardening. Also the correction proposed by Yoshida-Uemori for elastic modulus as a function of the equivalent plastic strain has been implemented.

The contents of this paper are as follows. The next section briefly describes the numerical approximation employed. Section 3, where the constitutive model used is presented, briefly describes the assumptions, the integration scheme of the constitutive equations and the correction equations for the elastic modulus due to plasticity. Section 4 shows numerical examples: uniaxial loading of a prismatic element and the deep drawing of a sheet to obtain a U-shaped channel. Finally, Section 5 summarizes the conclusions obtained in this work.

2 NUMERICAL APPROXIMATION

The numerical approach used is based on an eight-node solid element [7], where logarithmic strains are obtained from the spectral decomposition of the right Cauchy-Green tensor. The element has been developed to simulate nonlinear materials in large strains where it is necessary to include various elements across the sheet thickness. Therefore, it only aims to alleviate the problem of volumetric locking (when plastic flow problems are addressed) and transverse shear dominated by bending problems. With this aim an ANS is used for the calculation of transverse shear components of the metric tensor.

The element is simple, does not require any stabilization, and does not show instabilities in large strains, showing suitable for contact problems. A four-point integration is considered that makes it slightly more flexible which is an advantage in springback simulation.

Since the element is implemented in a code with explicit integration of the governing equations, the approach includes a strategy of selective mass scaling which allows that the critical time does not depend on the discretization across the thickness. The basic idea behind this strategy is to distinguish between the accelerations of the middle surface, i.e. the set of nodes on a regular fiber to the sheet, and the accelerations relative to the middle surface accelerations.

3 CONSTITUTIVE MATERIAL MODEL

This section briefly describes the model of Yoshida-Uemori [15], which is developed in the framework of a set of two surfaces. An inner yield surface (constant radius) moves with kinematic hardening within a bounding surface which shows mixed hardening (isotropic and kinematic). An illustrative diagram is shown in Figure 1. Both surfaces can be treated using Huber-Mises functions or Hill’48 type. The model reproduces the Bauschinger effect as other models of two surfaces do, but can also take into account the hardening stagnation that is experimentally observed in some materials under cyclic loading. The implicit scheme or the Yoshida-Uemori integration algorithm proposed in this paper is an alternative to that presented in [16].
3.1 Two-surface model description

The Yoshida-Uemori model [15] assume small elastic deformations and large plastic deformations. The strain rate tensor \( D \) can be decomposed as:

\[
D = D^e + D^p
\]

where \( D^e \) are the elastic and plastic parts of the strain rate tensor respectively.

3.1.1 Yield function and flow rule for model

The yield function used, similar to that proposed by Yoshida-Uemori is:

\[
f = \sqrt{\frac{2}{3}} (s - \alpha) N (s - \alpha) - Y = 0
\]

in (2) \( s \) is the deviatoric component of the stress tensor, \( \alpha \) is the back stress, and \( Y \) the material yield stress. For the sake of completeness the matrix with Hill coefficients (Hill, 1948) \( N \) has been included, unnecessary if Huber-Mises is considered as it is the identity matrix. This yield function is associated with the flow rule:

\[
D^p = \dot{\varepsilon}^p = \dot{\lambda} \frac{\partial f}{\partial s} = \frac{3}{2} \dot{\lambda} n^p
\]

where \( n^p = N (s - \alpha) / \| s - \alpha \| \) is the plastic flow direction and \( \dot{\varepsilon}^p = \sqrt{\frac{2}{3}} \dot{\varepsilon}^p : \dot{\varepsilon}^p = \dot{\lambda} \) is the effective plastic strain rate.

The bounding surface is defined by a function of the type:
\[ F = \sqrt{\frac{1}{2} (s - \beta) N (s - \beta)} - (B + R) = 0 \]  \hspace{1cm} (4)

where \( \beta \) is the back stress of the limit surface, while \( B \) and \( R \) are the initial radius and the isotropic hardening component of the bounding surface respectively.

### 3.1.2 State variables evolution for the model

The Yoshida-Uemori model\[15\] proposes a kinematic evolution of the yield surface \( f(s, \alpha) \) and a mixed evolution of the bounding surface \( F(s, \beta, \varepsilon^p) \). The relative movement between both surfaces is established using the difference in back stress measures \( \alpha_* = \alpha - \beta \) (see Figure 1), whose evolution is:

\[ \dot{\alpha}_* = C \left( \frac{2}{3} b \dot{\varepsilon}^p - \sqrt{\frac{a}{\bar{\alpha}}} \alpha_* \dot{\varepsilon}^p \right) = Ca \left[ \mathbf{n}^p - \sqrt{\frac{a}{\bar{\alpha}}} \mathbf{n}_\alpha \right] \dot{\lambda} \]  \hspace{1cm} (5)

where \( \dot{\varepsilon}^p \) is the change in effective plastic strain \( (\dot{\varepsilon}^p = \dot{\lambda} \) according to (3)), \( \mathbf{n}^p \) and \( \mathbf{n}_\alpha \) denote the unit vectors along the directions of \( \mathbf{D}^p \) and \( \alpha_* \), and \( C \) is a material parameter which regulates the kinematic hardening rate of the yield function. Since the yield function must keep inside the bounding surface, a condition of the type \( \|\alpha_*\| \leq a = B + R - Y \) emerges that defines parameter \( a \). In the case of anisotropic plasticity (Hill'48) \( \mathbf{n}_\alpha = \mathbf{N} \alpha \|\alpha\| \) is adopted.

The bounding surface presents two evolution equations, the isotropic hardening evolution is:

\[ \dot{R} = m \langle R_{sat} - R \rangle \dot{\varepsilon}^p = m \langle R_{sat} - R \rangle \dot{\lambda} \]  \hspace{1cm} (6)

where \( R_{sat} \) is the saturation stress of the isotropic hardening as the plastic strain is large enough. Furthermore \( m \) is the parameter that regulates the isotropic hardening rate of the bounding surface.

For the kinematic hardening of the bounding surface, the evolution of the stress \( \beta \) results:

\[ \dot{\beta} = m \left( \frac{2}{3} b \dot{\varepsilon}^p - \beta \dot{\varepsilon}^p \right) = mb \left[ \mathbf{n}^p - \left( \frac{\bar{\beta}}{b} \right) \mathbf{n}_\beta \right] \dot{\lambda} \]  \hspace{1cm} (7)

where \( b \) is a material parameter and \( m \) is the same value adopted in (6). Note that as in (5), the evolution equation (7) has been written in terms of the directions \( \mathbf{n}^p \) and \( \mathbf{n}_\beta \) corresponding to tensors \( \mathbf{D}^p \) and \( \beta \). Being \( \bar{\beta} = ||\beta|| \) the bounding surface back stress modulus while for anisotropic plasticity \( \mathbf{n}_\beta = \mathbf{N} \beta / ||\beta|| \).
3.1.3 Hardening stagnation of the bounding surface

The experimental evidence suggests that some materials show a hardening stagnation for cyclic deformation [14, 15]. This can be described as a non-isotropic hardening of the bounding surface $F$, i.e. only the kinematic component in (4) is non zero. The non-isotropic hardening surface $g_\sigma$ results:

$$g_\sigma(\beta, q, r) = \sqrt{\frac{3}{2}} (\beta - q) : (\beta - q) - r$$

where $q$ is the stress that defines the center of the surface and $r$ is its radius. The isotropic hardening of the bounding surface $F$ (i.e. the condition $\dot{R} > 0$), occurs when the central point of this surface (defined by $\beta$) is on the stagnation surface $g_\sigma(\beta, q, r) = 0$ and also $(\partial g_\sigma/\partial \beta) : \dot{\beta} > 0$.

In this work it is assumed that the center of the surface $g_\sigma$ moves in direction $(\beta - q)$ as:

$$\dot{q} = \mu (\beta - q)$$

while the surface radius evolution follows:

$$\begin{cases}
\dot{r} = h \Gamma = h \left[ \frac{3(\beta-q):\beta}{2r} \right] \rightarrow \dot{R} = m < R_{sat} - R > \dot{\varepsilon} > 0 \\
\dot{r} = 0 \rightarrow \dot{R} = 0
\end{cases}$$

3.2 Constitutive equation integration

A Newton-Raphson scheme has been used to satisfy (2), so when the trial stresses lead to a positive value of the yield function, it is raised:

$$f^{tr} + \frac{df}{d\lambda} \Delta \lambda = f^{tr} + \left[ \frac{df}{d\lambda} \frac{d\alpha_{*}}{d\alpha} + \frac{df}{d\alpha} \left( \frac{d\alpha_{*}}{d\lambda} + \frac{d\beta}{d\lambda} \right) \right] \Delta \lambda = 0$$

it can be seen that in (11) all the terms can be obtained from the equations described above, and considering that $s = s^{tr} - 2\mu \lambda n^p$. Working on the equation (11) the increment in the consistency parameter is obtained:

$$\Delta \lambda = \frac{f^{tr}}{(2G + Ca + mb) n^p : n^p - (C\sqrt{a\alpha n^p : n_{*}} + m\beta n^p : n_{*})}$$

initially, when $\alpha_{*} = 0$ and $\beta = 0$, a radial return is obtained.

Then, considering (9), (10) and that initially $r = 0$ and $q = \beta$, the change in the stagnation function implies $\frac{d}{dt} g_\sigma = 0$ so the increment in the stagnation consistency parameter can be computed as:
\[ \Delta \mu = 1 - \left( \frac{r + \Delta r}{r^{tr}} \right) \]  

(13)

From the value of \( \Delta \lambda \) obtained in (12) the state variables of the model can be updated:

\[ s_{n+1} = s_{n+1}^{tr} - 2\mu\Delta \lambda n^p \]  

(14)

\[ \alpha_{n+1} = \alpha_{n} + C a_n \left[ n^p - \sqrt{n_{*}^p} n_{*} \right] \Delta \lambda \]  

(15)

\[ \beta_{n+1} = \beta_{n} + m b \left[ n^p - \left( \frac{e_n^p}{n_{*}} \right) n_{*} \right] \Delta \lambda \]  

(16)

the back stress of the yield function \( \alpha_{n+1} = \alpha_{n} + \beta_{n+1} \) and the effective plastic strain \( e_{n+1}^p = e_n^p + \Delta \lambda \) can also be updated.

Finally, and only for the case \( g_{\sigma} (\beta_{n+1}, q_n, r_n) = 0 \), the isotropic component of the bounding surfaces is updated:

\[ R_{n+1} = R_n + m < R_{sat} - R_n > \Delta \lambda \]  

(17)

and using \( \Delta \mu \) the stagnation variables:

\[ q_{n+1} = q_n + \Delta \mu \left( \beta_{n+1} - q_n \right) \]  

(18)

\[ r_{n+1} = r_n + h \left\{ 3 \left( \beta_{n+1} - q_n \right) : \Delta \beta \right\} / 2r_n \]  

(19)

3.3 Elastic modulus correction in large strains

According to [15] the evolution of the elastic modulus as a function of the plastic strain is:

\[ E = E_0 - \left( E_0 - E_0 \right) \cdot \left( 1 - \exp \left( -\xi e^p \right) \right) \]  

(20)

in (20) \( E_0 \) represents the initial elastic modulus , \( E_0 \) is the elastic modulus of the material for infinite strain, and \( \xi \) is the exponent that describes the change in elastic modulus. The variable that rules the evolution of elastic modulus is the effective plastic strain \( e^p \).

4 NUMERICAL EXAMPLES

Below the results of two separate analyzes are presented. First a specimen under unidirectional cyclic loading condition has been studied. Then the results for the deep drawing of a sheet are presented. The results obtained with the formulation proposed in this work are identified with the acronym SOLAG.
4.1 Uniaxial loading case analysis

The uniaxial cyclic loading analysis is carried out considering a specimen with the dimensions proposed in the work [14]. The sample has a cross section of 5mm thick and 10 mm wide, while the effective length is 14 mm. This geometry has been meshed with a single element.

- First, CASE A, a high strength steel has been considered, with mechanical properties defined by: \( Y = 143 \text{ MPa}, \ C = 557, \ B = 280 \text{ MPa}, \ R_{sat} = 147 \text{ MPa}, \ b = 23 \text{ MPa}, \ m = 29 \) and \( h = 0.5 \) for the plastic behavior, and Young modulus \( E = 206 \text{ GPa}, \ E_a = 160 \text{ GPa} \) and \( \xi = 60 \) for the dependance of the elastic parameters with the plastic strain

- Second, CASE B, a mild steel with plastic properties: \( Y = 124 \text{ MPa}, \ C = 500, \ B = 168 \text{ MPa}, \ R_{sat} = 190 \text{ MPa}, \ b = 9 \text{ MPa}, \ m = 12 \) and \( h = 0.5 \), and elastic properties defined by Young modulus \( E = 206 \text{ GPa}, \ E_a = 152 \text{ GPa} \) and \( \xi = 30.8 \).

For CASE A a cyclic strain \( \varepsilon = \pm 0.02 \) is considered, while for CASE B the cyclic strain is in the range \( \varepsilon = 0.00 - 0.02 \). The results obtained are those presented in Figure 2, that are compared with those obtained by Yoshida.

For CASE A a cyclic strain \( \varepsilon = \pm 0.02 \) is considered, while for CASE B the cyclic strain is in the range \( \varepsilon = 0.00 - 0.02 \). The results obtained are those presented in Figure 2, that are compared with those obtained by Yoshida.

![Figure 2: Results for the uniaxial cyclic test simulation. (a) CASE A. (b) CASE B.](image-url)

The results show significant similarity to those obtained by Yoshida ([15] and [17]), even more if it is considered that the mesh includes just one 8-node solid element. Although the results are in closer agreement for the mild steel (CASE B), in the CASE A (high strength
steel) the constitutive model used more clearly shows the effect of hardening stagnation when load is reversed (which should be easily appreciated considering the adopted value for \( h \) and the larger capability to develop hardening due to the value of \( m \)).

4.2 Springback in a channel draw

This problem corresponds to the example number three (BM3) of Numisheet’05. It is the deep drawing of a U-shaped channel starting from a 1mm thick flat sheet made of a dual phase steel DP600. The analysis is performed until the deformed piece is extracted and the elastic recovery is measured. A mesh with 8040 eight-node solid elements is used where only three elements across the thickness are considered.

The figure 3 shows the dimensions and geometry of the die, the punch, the drawbead and the kiss blocks. The kiss-blocks are used to keep a constant clearance of 1.42 mm between the blank-holder and the die. The load on the blank-holder is 637 kN, and the friction coefficient of \( \mu = 0.15 \) between the blank and the tools has been adopted for the simulation.

Two material models have been considered: (a) the two-surface model of Yoshida-Uemori with plastic properties \( Y = 420 \text{ MPa}, C = 200, B = 555 \text{ MPa}, R_{sat} = 190 \text{ MPa}, b = 110 \text{ MPa}, m = 12 \) and \( h = 0.9 \), and (b) an elastic-plastic model with saturation of the Voce type \( \sigma (\varepsilon) = 424 \text{ MPa} + 387 \text{ MPa} (1 - \exp(-25\varepsilon)) \). For both constitutive models plastic anisotropy has been considered with Lankford parameters \( r_{0} = 0.821, r_{45} = 0.915 \) and \( r_{90} = 0.969 \). Besides, the elastic properties adopted for the material are \( E = 206 \text{ GPa} \) and \( \nu = 0.30 \), while a modulus of \( E_a = 163 \text{ GPa} \) for infinite strain and an exponent \( \xi = 135 \) for the elastic modulus correction due to plastic strains were adopted.
Figure 4: Deformed sheet after springback. (a) Without correction in elastic modulus due to plasticity. (b) Considering reduced elastic modulus due to plasticity.

The Figure 4 shows the computed deformed sheet profiles after springback and those presented in [18] where a plane strain model with a reduced integration element has been used. It can be observed that the use of reduced elastic modulus considerably improves the results for the two-surface Uemori-Yoshida model. Something similar is observed for the Voce isotropic hardening model when it is used together with equation (20). The discrepancies between the results obtained in this study with those observed in [18], may be due to the different approaches and possibly to the coarse across the thickness discretization used here. The Figure 5 shows the deformed geometry of the sheet after springback, and a detail of the thickness discretization is also observed.

Figure 5: Deformed blank after springback for Yoshida-Uemori model. (a) Without correction in elastic modulus due to plasticity. (b) Considering reduced elastic modulus due to plasticity.
5 CONCLUSIONS

In this paper a two-surface plasticity model proposed by Yoshida-Uemori has been implemented, which allows to consider both isotropic and kinematic hardening material. This constitutive model is used together with on a solid element with only four Gauss points and an assumed natural strain for transverse shear components to improve the bending behavior of the element.

The proposed solution scheme shows appropriate behavior when it is applied to plastic forming of thin sheets. From the standpoint of the springback behavior is more rigid compared to that provided by shell elements. However this could be improved using an adequate discretization across the thickness.

With regard to CPU time analysis of these type of problems, it can be reduced by applying selective mass scaling. Although this scheme was available it was not used for the simulation of the problems studied here.

ACKNOWLEDGMENTS

The authors gratefully acknowledge financial support from CONICET and SeCyT-UNC.

References


A FFT BASED MESOSCOPIC APPROACH FOR THE COMPRESSION AND RECOVERY SIMULATION OF STRUCTURED NONWOVENs

SARAH STAUB, HEIKO ANDRÄ AND MATTHIAS KABEL

Fraunhofer ITWM
Fraunhofer-Platz 1
67663 Kaiserslautern
Germany
sarah.staub@itwm.fraunhofer.de
heiko.andrae@itwm.fraunhofer.de
matthias.kabel@itwm.fraunhofer.de
http://www.itwm.fraunhofer.de

Key words: FFT, Nonwovens, Compressional Behavior, Viscoelastic Recovery

Abstract. In the work at hand a three-dimensional mesoscopic model for the compression and recovery simulation of structured nonwovens is presented. Starting point of the model is a one-dimensional power-law, which is extended towards a three-dimensional orthotropic model incorporating the local fiber volume fraction. Furthermore, strain rate dependency of fibrous materials is considered. The presented constitutive model is implemented into the solver FeelMath, which is using the fast Fourier transform to solve the Lippmann-Schwinger equations.

1 INTRODUCTION

The industrial manufacturing of components requires the forming and shaping of porous materials by e.g., embossing, perforation, pleating, and compression. This macroscopic structuring is necessary to fit the media into the design space and to improve the functionality (e.g. pleated filter media). In order to simulate a structure as presented in Figure 1 on a fiber network level complicated contact problems, where frictional contact between single fibers have to be taken into account. On the other hand if the structure is considered on a macroscopic scale as a homogeneous continuum the influence of the strong fluctuation of the fiber densities is neglected. In the current work we propose a meso-scale model which is a compromise between these two approaches. The local fiber densities are obtained based on the analysis of highly resolved μCT images and transferred to a continuum scale, see Figure 2.
In the current approach for the virtual design of structured nonwovens the software GeoDict \cite{1} can be used to generate fiber networks. In the next step a virtual embossing and a coarsening leads to the mesoscopic geometry, which can be easily analyzed by the proposed method.

## 2 GOVERNING EQUATIONS

In the following, the governing equations for the boundary value problem (BVP), its reformulation in terms of the Lippmann-Schwinger (LS) equation and its solution in the Fourier space are outlined.

For the applicability of the fast Fourier transform (FFT) we consider a rectangular domain $\Omega$ on which a constant strain $E$ is applied. The unknown displacement field $u = E \cdot x + u^*$ fulfills the equilibrium condition,

$$\text{div} \, \sigma(x) = 0, \quad x \in \Omega,$$

in terms of the Cauchy stress $\sigma$. Please note that here the influence of volume forces is omitted. The relation between the stresses and the strains is given by the constitutive equation

$$\sigma(x) = C(x) : \epsilon(x), \quad x \in \Omega. \tag{2}$$

Herein, $C(x)$ corresponds to the local (anisotropic) stiffness tensor. The kinematics of the BVP read

$$\epsilon(x) = E + \frac{1}{2} \left[ \text{grad} u^*(x) + \text{grad}' u^*(x) \right], \quad x \in \Omega. \tag{3}$$

The governing equations are completed by the periodic displacement boundary conditions $u^*(x)$ on $\partial \Omega_D$ and the anti-periodic tractions $t = \sigma(x) \cdot n$ on $\partial \Omega_N$.

### 2.1 Fast Fourier Transformation of Lippmann-Schwinger Equations in Elasticity

In a first step the differential equation (1) is reformulated according to Zeller and Dedereichs \cite{2} as the so-called Lippmann-Schwinger equation, see also Lippmann and Schwinger \cite{3}. To this end a constant homogeneous reference stiffness tensor $C^0$ is introduced and the polarization stress with respect to the reference stiffness reads

$$\tau(x) = \sigma(x) - C^0 : \epsilon(x). \tag{4}$$

The solution of the equilibrium equation $\text{div} \, \sigma = 0$ can be expressed by the nonlocal Green operator $\Gamma^0$ associated with the reference stiffness as

$$\epsilon(x) = E - (\Gamma^0 * \tau) (x). \tag{5}$$

Therein, the convolution is defined by

$$(\Gamma^0 * \tau)(x) = \int_{\Omega} \Gamma^0(x - y) : \tau(y) \, dy. \tag{6}$$
Finally, the LS equation is obtained from (5) and (4) as
\[
E = \epsilon(x) + \Gamma^0 \ast \left( [C(x) - C^0] : \epsilon(x) \right) \quad \text{(7)}
\]
\[
= (I + B \epsilon) \epsilon. \quad \text{(8)}
\]

Please note, that the Green operator $\Gamma^0$ is independent of the fluctuations and thus only depends on the homogeneous linear elastic reference stiffness and the boundary conditions, see Kröner [4].

The LS equation can be solved iteratively using the Neumann series expansion for inverting the operator $(I + B \epsilon)$. The iterates of the local strains read
\[
\epsilon^0 = E \quad \text{(9)}
\]
\[
\epsilon^{n+1} = -B \epsilon^n + E. \quad \text{(10)}
\]

These iterates can be efficiently calculated in four simple steps using FFT by the so-called basic scheme, see Moulinec and Suquet [5]:
\[
\tau = C : \epsilon^n - C^0 : \epsilon^n, \quad \text{(11)}
\]
\[
\hat{\tau} = \text{FFT}(\tau), \quad \text{(12)}
\]
\[
\hat{\eta} = -\hat{\Gamma}^0 : \hat{\tau}, \quad \hat{\eta}(0) = E, \quad \text{(13)}
\]
\[
\epsilon^{n+1} = \text{FFT}^{-1}(\hat{\eta}). \quad \text{(14)}
\]

Explicit expressions for the Green operator $\hat{\Gamma}^0$ in the Fourier space can be found in Mura [6], an overview of different solution schemes is given in Kabel et al. [7].

### 2.2 Elastic Constitutive Equations depending on Local Fiber Volume Fraction

Due to the fiber network, which is the basis of the considered structured nonwoven, an orthotropic model is developed for the simulation of these structures. The constitutive equation, see equation (2), can be rewritten in Voigt notation as
\[
\sigma^V = C^V \cdot \epsilon^V. \quad \text{(15)}
\]

Generally, for an orthotropic material the Voigt notation of the compliance matrix $S^V$, which is the inverse of the stiffness $C^V$, reads
\[
S^V = \begin{bmatrix}
\frac{1}{E_x} & -\nu_{y,x} & 0 & 0 & 0 \\
\nu_{y,x} & \frac{1}{E_y} & 0 & 0 & 0 \\
\frac{1}{E_{yz}} & -\nu_{z,x} & \frac{1}{E_x} & 0 & 0 \\
\frac{1}{E_{zx}} & \nu_{z,y} & 0 & \frac{1}{G_{yz}} & 0 \\
\text{sym} & 0 & 0 & 0 & \frac{1}{G_{xy}}
\end{bmatrix}. \quad \text{(16)}
\]
In this term $E_x$, $E_y$ and $E_z$ correspond to the Young’s moduli in the directions $x$, $y$, and $z$. Furthermore, $\nu_{ij}$ and $G_{ij}$, $i, j = x, y, z$ denote the Poisson’s ratios and the shear moduli in the corresponding directions.

Due to the fact that only compression simulations in thickness direction, e.g. $z$-direction, are considered in the following, the constitutive law in the other directions is assumed to be basically linear elastic. In thickness direction a fiber volume fraction (FVF) dependent power-law as proposed by van Wyk [8] is introduced. Within this (one dimensional) power-law the compressional stress $\sigma_c$ in thickness direction is connected to the fiber volume fraction $\Phi$ via

$$\sigma_c = a \Phi^b.$$  \hfill (17)

Therein, $a$ and $b$ denote material specific parameters which need to be fitted to compression experiments. The actual FVF can be determined depending on the area weight $A_f$ of the nonwoven, the actual height $L$ and the fiber density $\rho_0$ via

$$\Phi = \frac{A_f}{\rho_0 L}.$$  \hfill (18)

Finally, the relationship $\epsilon_c = \frac{L - L_0}{L_0}$, whereby $L_0$ denotes the initial thickness of the network, yields the desired one-dimensional connection between the stresses and strains as

$$\sigma_c = a \left[ \frac{A_f}{\rho_0 L_0 [1 + \epsilon_c]} \right]^b.$$  \hfill (19)

In a next step, the developed one-dimensional power-law is extended such that it can be incorporated into the formulation of the orthotropic compliance matrix, see (16). In a three-dimensional formulation, where the $z$-direction is considered as thickness direction, equation (19) reads

$$\sigma^V_3 = a \left[ \frac{A_f}{\rho_0 L_0 [1 + \epsilon^V_3]} \right]^b.$$  \hfill (20)

The corresponding entry of the compliance matrix is obtained through the following derivative

$$S^V_{33} = \frac{\partial \epsilon^V_3}{\partial \sigma^V_3} = \frac{A_f}{\rho_0 L_0 b a} \left[ \sigma^V_3 \frac{a}{b} \right]^{\frac{1}{b+1}}.$$  \hfill (21)

In order to ensure a numerically stable three-dimensional formulation a dependency on the FVF is also introduced in the other non-zero entries of $S^V$. For the diagonal values $S^V_{ii}, i = 1, 2, 3$, it is assumed that the values tend towards $1/E^*$ for FVF $\to 1$, where, $E^*$ denotes the Young’s modulus of a solid block of the polymer material of which the fibers are made from. Similarly, for the diagonal values $S^V_{ii}, i = 4, 5, 6$, it is assumed the values tend towards $1/G^*$ for FVF $\to 1$. The formulation for the off-diagonal values of $S^V$ is then straight-forward.
2.3 Extension of Constitutive Equation towards Viscoelasticity

In order to incorporate viscoelastic effects into the simulation, the orthotropic FVF dependent model from section 2.2 is extended by a general Maxwell model. In the following the viscoelastic model is outlined briefly. For a more detailed formulation including the theoretical background the reader is referred to Kaliske and Rothert [9]. In the considered viscoelastic model the stresses and the strains are related via the following convolution integral

\[ \sigma(x, t) = \int_0^t \Pi'(x, t - s) \frac{\partial \epsilon(x, s)}{\partial s} \, ds. \]  

(22)

Therein, \( \Pi' \) corresponds to the fourth-order relaxation tensor of the viscoelastic material. The viscoelasticity is realized by means of a generalized Maxwell model consisting of \( N \) Maxwell elements. In terms of the normalized relaxation coefficients \( \gamma_j, j = 1, ..., N \), and the relaxation times \( \tau_j, j = 1, ..., N \) the normalized relaxation function reads

\[ \Psi(t) = 1 + \sum_{j=1}^{N} \gamma_j \exp \left( -\frac{t}{\tau_j} \right). \]  

(23)

In the work at hand it is assumed that all components of the stiffness tensor have the same relaxation behavior and thus

\[ \Pi(t) = \Psi(t) C_r = C_r + \sum_{j=1}^{N} \gamma_j \exp \left( -\frac{t}{\tau_j} \right) C_r \]  

(24)

is valid. In this equation \( C_r \) corresponds to the relaxed fourth-order stiffness tensor, which here is assumed to be orthotropic and which depends on the FVF. It can be obtained by inverting the compliance \( S \) as derived in section 2.2.

The basic steps of the numerical implementation of the viscoelastic part of the constitutive model are given in the following. In a first step the elastic stresses \( \sigma_0 \) and the internal variables \( h_j^0, j = 1, ..., N \), are initialized, i.e.

\[ \sigma_0^0 = \sigma_r^0 \quad \text{and} \quad h_j^0 = h_j^*. \]  

(25)

Then in each discrete timestep \( \Delta t^{n+1} \) the elastic stresses are calculated via

\[ \sigma_0^{n+1} = C_r : \epsilon^{n+1}, \]  

(26)

and the stress-like internal variables are updated via

\[ h_j^{n+1} = \exp \left( -\frac{\Delta t^{n+1}}{\tau_j} \right) h_j^n + \gamma_j \frac{1 - \exp \left( -\frac{\Delta t^{n+1}}{\tau_j} \right)}{\frac{\Delta t^{n+1}}{\tau_j}} \left[ \sigma_0^{n+1} - \sigma_0^n \right]. \]  

(27)
Finally, the total stresses are obtained as

\[ \sigma^{n+1} = \sigma_0^{n+1} + \sum_{j=1}^{N} h_j^{n+1}, \]  

(28)

and the viscoelastic stiffness tensor reads

\[ C^{\text{vel,n+1}} = \Psi^{n+1} C' \quad \text{with} \quad \Psi^{n+1} = 1 + \sum_{j=1}^{N} \gamma_j \left( 1 - \exp \left( \frac{\Delta t^{n}}{\tau_j} \right) \right). \]  

(29)

3 GENERATION OF MESOSCOPIC STRUCTURES

In order to take the local fiber volume fraction of the structured nonwovens into account a so-called coarsening procedure is developed. This procedure has two tasks. Firstly, segmented \( \mu \)CT images, which resolve the fibers as well as the structuring of the nonwoven, are transformed into an image containing the local FVF, which then can be used as an input for the constitutive laws as given in Section 2. Secondly, after the coarsening the regarded structures are captured in a coarser mesh and thus the number of degrees of freedom in the simulation is reduced drastically. Within the coarsening procedure several voxels of the input image are merged to one bigger voxel and the local FVF in this new voxel is obtained by averaging the binary values of the accordant input voxels. The number of voxels which are merged in one direction corresponds to the coarsening factor.

Alternatively, to \( \mu \)CT images, virtual structures can be considered to study the influence of different microscopic and morphological parameters onto the effective behavior. In Figure 1 an example of such a structured nonwoven is displayed. Here, the virtual

![Figure 1: Virtual structured nonwoven at microscale](image-url)
plain nonwoven is generated with the software package GeoDict [1]. The structuring has been applied in a postprocessing step. In Figure 2 the result of the coarsening procedure applied to the structure shown in Figure 1 is compared for different coarsening factors between 8 and 64. The microstructure used as input consists of 1200x1200x672 voxel with a resolution of 10 $\mu$m. Thus, the image coarsened with factor 8 consists of 150x150x84 voxel with a resolution of 80 $\mu$m.

4 NUMERICAL EXAMPLES

In this section two numerical examples are presented. The simulations of the viscoelastic model depend on the local FVF are carried out in the LS-FFT based solver FeelMath, which is developed at Fraunhofer ITWM. On the microstructure the geometry displayed in Figure 1 is considered. On top and on bottom of the structure a rigid plate is added, see Figure 3. In the following examples the coarsening factor 16 is used, i.e. the considered geometry has a size of 75x75x52 voxels. For the viscoelastic contribution of the model two Maxwell elements with relaxation times $\tau_1 = 2$ s and $\tau_2 = 6$ s are assumed.

In a first example a relaxation test for compression in thickness direction is considered. To this end a compressional strain of 5% is applied to the structured nonwoven at different strain rates and then held constant, see Figure 4. The strains are applied at strain rates of 0.05/s and 0.005/s. The resulting stresses in thickness direction are displayed on the right-hand side of Figure 4. Larger stresses are obtained for the faster strain rate, compare

**Figure 2:** Comparison of local FVF for different coarsening factors – a) 8, b) 16, c) 32, d) 64
Figure 3: Coarsened structured nonwoven with rigid plate at top and bottom

Figure 4: a) Applied compressional strains in thickness direction at 0.05/s and 0.005/s, b) resulting stresses in thickness direction

also Kelly et al. [10] Furthermore, the relaxation of the stresses is observed. Thus, the developed model is suitable to simulate the viscoelastic behavior of structured nonwovens including the local FVF.

In a second example a compression recovery test on the same geometry is considered. The applied strains are given at the left-hand side of Figure 5. The resulting stresses are given at the right-hand side of Figure 5. The stresses in thickness direction show the typical compression recovery behavior of nonwovens, compare e.g., Das and Pourdeyhimi [11]. After unloading a residual stress of -5 kPa remains in the structure. This effect is due to the relative long relaxation times of the Maxwell elements compared to the loading and unloading time.

5 Conclusion

A viscoelastic constitutive model for the compression simulation of structured nonwovens at a mesoscopic scale has been developed. With assistance of the coarsening procedure the local FVF can be incorporated into the model. The simulation via solving
the Lippmann-Schwinger equations within the solver FeelMath allows a fast and efficient simulation of complex structures. In numerical examples it has been shown that relaxation tests as well as compression recovery tests yield the typical behavior of nonwovens as given in the literature. Thus, in the future work the influence of the properties of the plain nonwoven as well as the structuring pattern onto the effective compression and recovery behavior can be analyzed.

To sum up, the developed method yields a promising compromise between precision and efficiency, which can be adaptively controlled in terms of the coarsening factor.

REFERENCES


VALIDATION OF MULTISCALE MODEL FOR HEAT GENERATION IN HARDENING CONCRETE

WILSON R. LEAL DA SILVA, VÍT ŠMILAUER AND ZDENĚK BITTNAR

1 Czech Technical University in Prague, Faculty of Civil Engineering, Dept. of Mechanics
Thákurova 7, 166 29 Prague 6, Czech Republic
* Corresponding author: bittnar@fsv.cvut.cz | http://mech.fsv.cvut.cz/~bittnar

Key words: hydration heat, FEM, mass concrete, multiscale model, temperature

Abstract. Urbanization has increased dramatically the demand of concrete production and quality control in developing countries. At present, several mass concrete elements are under construction in Brazil and India, to mention a few; and it is important to assure that the thermal behavior of such structures complies with local regulations to avoid durability issues. Our approach to predict the thermal behavior of mass concrete is based on a semi-adiabatic calorimeter setup and FE analysis. This combination provides evolution of hydration heat under isothermal temperature and enables upscaling to the temperature evolution in mass concrete on a multi-scale level. The upscaling approach is demonstrated on a 511m³ mass concrete block.

1 INTRODUCTION

The concrete industry in developing countries has observed an increased demand for mass concrete structures. The demand for durability of such structures has increased likewise. Mass concrete structures are prone to early age cracking due to thermal stresses. The prediction of temperature rise and its gradients is a classical approach for crack mitigation.

Our approach for predicting concrete temperature relies on the affinity hydration model. [1, 2] The parameters of this model can be calibrated from isothermal calorimetry. However, semi-adiabatic measurements are used more extensively due to its simplicity. In this case, temperature evolution depends on concrete thermal properties and boundary conditions; thus, Finite Element (FE) simulation is needed to account for non-homogeneous temperature across a sample.

The objective of this work is to validate the upscaling approach for heat generation in hardening concrete that is illustrated in Fig. 1. For that, semi-adiabatic measurements and FEM analysis are combined to calibrate the affinity hydration model and simulate the thermal behavior of hydrating concrete at different length scales, Simulation results are validated on a 511 m³ mass concrete block cast in Southern Brazil.
2 Affinity hydration model and heat transport formulation

The affinity hydration model provides a framework for accommodating all stages of cement hydration under isothermal temperature at 25°C. The rate of hydration can be expressed by

\[
\frac{d\text{DoH}}{dt} = \tilde{A}_{25}(\text{DoH}),
\]

where \(\tilde{A}_{25}(\text{DoH})\) is the chemical affinity. In this work, a modified formulation is taken into account \([1-3]\)

\[
\tilde{A}_{25}(\text{DoH}) = \beta_1 \left( \frac{\beta_2}{\text{DoH}_\infty} + \text{DoH} \right) (\text{DoH}_\infty - \text{DoH}) \cdot \exp \left( -\eta \frac{\text{DoH}}{\text{DoH}_\infty} \right),
\]

(1)

where \(\beta_1\) and \(\beta_2\) are empirical parameters to be adjusted, \(\eta\) the micro-diffusion of free water through formed hydrates, and \(\text{DoH}_\infty\) the ultimate degree of hydration. When hydration proceeds under varying temperature, Eq. (1) can be scaled via Arrhenius equation, i.e.

\[
\tilde{A}_T = \tilde{A}_{25} \cdot \exp \left[ \frac{E_a}{R} \cdot \left( \frac{1}{298.15} - \frac{1}{T} \right) \right],
\]

(2)

where \(R\) [Jmol\(^{-1}\)K\(^{-1}\)] is the universal gas constant and \(E_a\) [Jmol\(^{-1}\)] is the activation energy. The affinity hydration model performs well on OPC and blended cements as indicated in Fig. 2 and published elsewhere. \([1, 4]\)

![Figure 2: Validation of the affinity hydration model on OPC and blended cements](image-url)
The transient three dimensional heat conduction problem in incompressible media is derived from a well-known energy balance on a differential element. The combination of this with Fourier's law yields

$$\lambda(x) \Delta T(x, t) + \overline{Q}(x, t) = \rho(x) c_V(x) \frac{\partial T(x, t)}{\partial t},$$

(3)

where $T(x, t)$ [K] is the unknown temperature field, $\lambda(x)$ [Wm$^{-1}$K$^{-1}$] the thermal conductivity of an isotropic material, $\rho(x)$ [kg/m$^3$] the material density, $c_V(x)$ [Jkg$^{-1}$K$^{-1}$] the specific heat capacity, and $\overline{Q}(x, t)$ [W/m$^3$] the concrete hydration heat. Initial and boundary conditions can be assigned to Eq. (3). A weak solution is obtained by multiplying Eq. (3) with a virtual temperature field.[5] After a short elaboration, a weak formulation reads

$$C \dot{r} + K r = p,$$

(4)

where $K$ is the conductivity matrix, $C$ the capacity matrix, and $p$ the heat load vector, which captures both the boundary conditions and the heat source.[6] A v-form version of the trapezoidal scheme is used for solution in the form of predictor-corrector.[7]

### 3 Application

The approach illustrated in Fig. 1 is demonstrated on a mass concrete block. First, semi-adiabatic measurements are used as a reference for calibrating cement hydration parameters. Next, the calibrated model is upscaled and validated on a 511 m$^3$ mass concrete block. The concrete mix design listed in Table 1 was used in both the semi-adiabatic setup and the mass concrete block. Concrete thermal conductivity and heat capacity were set constant at 1.8 Wm$^{-1}$K$^{-1}$ and 870 Jkg$^{-1}$K$^{-1}$, respectively, which are reasonable assumptions according to the literature and the concrete composition (Table 1).[8, 9]

<table>
<thead>
<tr>
<th>Table 1: Mass concrete mix design</th>
<th>kg/m$^3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Materials</td>
<td></td>
</tr>
<tr>
<td>Binder (OPC+Class F Fly ash)</td>
<td>420.0</td>
</tr>
<tr>
<td>Fine aggregate</td>
<td>750.0</td>
</tr>
<tr>
<td>Coarse aggregate</td>
<td>1020.0</td>
</tr>
<tr>
<td>Water</td>
<td>138.6</td>
</tr>
<tr>
<td>Crushed ice</td>
<td>59.4</td>
</tr>
<tr>
<td>Water-reducing admixture*</td>
<td>0.65%</td>
</tr>
<tr>
<td>High-range water reducing admixture*</td>
<td>0.44%</td>
</tr>
</tbody>
</table>

*Percent of binder mass

### 3.1 Calibration of cement hydration parameters

The semi-adiabatic experimental setup consists of a concrete cube with 240 mm edge surrounded with a 100 mm thick polystyrene foam (Fig. 3a). Concrete temperature is recorded by
a K-type thermocouple placed in the middle of the concrete block.

The geometry of the experimental setup was discretized into linear brick finite elements and transferred into open-source OOFEM.\textsuperscript{[10]} Initially, $Q_{\text{pot}}$ was set at 518.37 J/g; however, the used binder was a fly ash blended cement. It turned out from our simulations that 45% of the binder needs to be replaced with any supplementary material such as class F fly ash, which yields low heat contribution in the first days of hydration.\textsuperscript{[5]} Thus, the effective amount of clinker is of 230 kg/m\textsuperscript{3} of concrete. The initial concrete temperature was set at 24.5\textdegree C as per temperature measurements. Boundary conditions of the polystyrene insulation are detailed in Fig. 3b. Only one fourth of the block was modeled due to symmetry conditions.

Figure 3: Semi-adiabatic experimental setup: a) experimental setup, b) FEM model and temperature field at 15.5 h, and c) calibration of hydration parameters

Hydration parameters from Eqs. (1) and (2) were fitted to $\beta_1 = 0.0007$ s\textsuperscript{-1}, $\beta_2 = 6.0 \times 10^{-5}$, $\eta = 6.1$, $DoH_\infty = 0.85$, and $E_a = 38.3$ kJ/mol. The model calibration is displayed in Fig. 3c, which indicates that the hydration kinetics was fitted reasonably well. Fig. 3b displays the temperature field for a quarter of the cube at the temperature maximum, which reached 55.2\textdegree C at 15.5 h.

3.2 Validation on mass concrete block

The foundation block of a residential building with 35 floors is located in Southern Brazil. The block has the dimensions of 19.6 $\times$ 10.1 $\times$ 2.5 m (511.0 m\textsuperscript{3}), see An overview of the block is shown in Fig. 4a. Temperature at the top of the block was assumed constant as per ambient temperature measurements. The exposed concrete surface of the block was covered with a water layer to assure adequate curing after casting completion.

The continuous concrete casting was modeled as two concrete layers with 6 hours time offset, leading to a better temperature agreement in the upper layers. Details on block geometry, mesh, boundary conditions, and temperature gauges location are depicted in Fig. 4b. Only one fourth of the block was modeled due to symmetry conditions; specifically, 3840 brick finite elements created the geometry. The thermal properties and initial temperature of soil and concrete layers are detailed in Fig. 4b.
Fig. 4c validates the temperature evolution in the concrete block. Temperature gauges 2, 4, 6 gave similar temperature evolution, which was only slightly influenced by the boundary conditions. The maximum concrete temperature reached 65°C at 74 h of hydration time (Fig. 4d).

4 Conclusion

The described upscaling approach proved successful for predicting mass concrete temperature at different scales. In particular, the parameters of the affinity hydration model can be estimated based on a semi-adiabatic calorimetry, which is a low-cost alternative to the standard isothermal calorimetry.

The upscaling of laboratory experiments to large-scale structures features a low cost and pragmatic approach, representing an alternative solution for the concrete industry in developing countries. At present, a semi-automatic tool with graphical interface is being developed at the CTU in Prague on the basis of the approach described in this paper.
5 Acknowledgments

We gratefully acknowledge the support from the project "Support for improving teams in research and development and the development of intersectoral mobility at Czech Technical University in Prague" OP VK CZ.1.07/2.3.00/30.0034, which allowed for funding of Dr. Leal da Silva’s postdoctoral research, and from the Competence Centres of the Technology Agency of the Czech Republic (TACR) under the project "Centre for Effective and Sustainable Transport Infrastructure" - project No. TE01020168.

References


Computational material modeling of hydrated cement paste calcium silicate hydrate (C-S-H) chemistry structure — influence of magnesium ion exchange on mechanical stiffness: C-S-H Jennite

Ram V. Mohan¹, Wayne D. Hodo², Babatunde Adebiyi¹, John Rivas¹, and Ahmed Mohamed¹

¹Department of Nanoengineering, North Carolina A&T State University
2907 E Lee Street, Greensboro, NC 27401, USA
rvmohan@ncat.edu; http://jsnn.ncat.uncg.edu

²U.S. Army Engineer and Research Development Center
3907 Halls Ferry Road, Vicksburg, MS 39180, USA
Wayne.D.Hodo@usace.army.mil; http://www.erdc.usace.army.mil

Key words: Material Chemistry, Calcium Silicate Hydrate, Magnesium Ion Exchange, Mechanical Stiffness Modulus

Abstract. Cementitious materials have material, geometric features ranging from molecular/nano, micro-, meso- and macro scales. Features and changes in material chemistry/nano scale influence the hydration process, formed micro scale morphology, associated properties and behavior at engineering length scales. In the present paper, effect of material ion exchange on mechanical stiffness properties of nano level hydrated cement paste constituent - calcium silicate hydrate (C-S-H) based on its material chemistry structure are studied following a molecular dynamics (MD) computational modeling methodology. Calcium ions are replaced with Magnesium ions in traditional C-S-H Jennite crystal structure of the hydrated cement constituent. Traditional C-S-H Jennite structure, and Magnesium modified structures obtained by calcium exchange of magnesium are used in the MD based material modeling analysis. Extensive modeling analysis established optimal computational MD modeling parameters and molecular material sizes employed. The present study clearly shows the influence of material chemistry due to Magnesium ion exchange in CSH-Jennite with predicted elastic modulus that changes with the level of Magnesium in CSH-Jennite material chemistry structure.

1 INTRODUCTION

Cementitious materials have material, geometric features ranging from molecular/nano, micro-, meso- and macro scales. Features and changes in material chemistry/nano scale influence the hydration process, formed micro scale morphology, associated properties and behavior at engineering length scales. Cement paste is made by combining dry Portland cement and water. The process of making cement paste is referred to as hydration, during which, there is loss of workability, solidification and hardening. The hydration of the different clinker phases results in different hydration product phases, which makes up the cement paste.
The hydrated cement phases include; calcium silicate hydrate (C-S-H), calcium aluminum hydrate (C-A-H) and calcium hydroxide (C-H)\(^1,2\). In cement paste, four levels of scales have been identified\(^3\) i.e. molecular /nano-scale level determined by the material chemistry features, micro-scale with microstructural morphology features, meso-scale and macro-scale. Distinct length scales and features are based on the requirement that each scale should be separated from the next scale by at least one order of length magnitude, which is also stated in the literature as a prerequisite for the application of continuum micromechanics\(^1,2\). As a result of the multi-scale nature of cement paste, the properties and material composition of the macro and engineering scale cement can be controlled or affected by the molecular level. A potential ability to alter and manipulate the chemistry level structure of hydrated component of traditional C-S-H, a structure predominantly influenced by calcium - a major constituent of the cement paste at the fundamental material chemistry level - will imply that the elastic and deformation properties of the paste would be impacted. Figure 1 shows the different levels of scale associated with cementitious materials.

![Figure 1: Schematic of multi-scale structure of cement based materials \(^3\)](image)

Traditional cement paste is primarily calcium-based, but there is interest in the engineering community for alternate formulations of cement; which may provide tailored properties for specific applications, could utilize local, and cheaper materials. Another factor influencing the alternate forms cement cited in the literature is the need to reduce the energy required to produce the clinker phase, and the carbon footprint of the cement industry\(^1\). Hence there have been efforts to look for alternate cement formulations that are not calcium-based, might be cheaper, and have the potential for reducing carbon footprint in the manufacturing stage of calcium based cement. Several forms of cement formulations has been sought, for instance, other forms of belitic cements are sought by trial and error to improve the reactivity of the clinker phase. Some additional approaches that have been used trying to improve the properties of cement-based materials include thermal processing\(^4,5\) and addition of new chemical compounds and ions\(^4,5,6\). Most common types of ionic substitution which has been used in cement include; Mg\(^{2+}\) for Ca\(^{2+}\), 2Al\(^{3+}\) and 2Fe\(^{3+}\) for 3Ca\(^{2+}\), and 2Ca\(^{2+}\) for Si\(^{4+}\)\(^6,7\). The present work investigates the influence of replacement of calcium ion with magnesium ion in the material chemistry level of C-S-H with a focus on mechanical stiffness determined via
Several studies have focused on different aspects of cement material improvement. Hegoi and co-workers\cite{6} studied the effects of the presence of chemical substitution on the physicochemical properties of cement clinker phases - alite and belite. They incorporated Mg$^{2+}$, Al$^{3+}$ and Fe$^{3+}$ into the structure using classical forcefield methods and reported that the crystallographic site within the unit cell is equally probable for Mg, Al and Fe substitution. It has also been reported that Mg incorporation does not change the electronic structure appreciably compared to Al and Fe incorporation\cite{6}, and that there are no preferential substitution for any calcium sites, when Mg$^{2+}$, Al$^{3+}$ and Fe$^{3+}$ were incorporated. The main focus of their research was on the reactivity of the clinker phase for reduced energy production.

As discussed in the literature, Mg provides a feasible incorporation option in cement paste. In a study of the hydration process of cement, Stephan and co-workers\cite{7}, provided analytical evidences that in low concentrations, Mg$^{2+}$ does not change the hydration process of calcium trisilicate (C$_3$S) – a component of cement, although the reactivity may increase with weight percent and with cement age. Also, because of the relative availability and the ionic similarities to Calcium, Mg can be considered as a potential candidate for ionic replacement. Additionally, with the replacement of Ca by Mg, a lower temperature for processing the clinker phase can be achieved.

The focus of this paper is to understand the influence of the effects of the exchange of Calcium ion with Magnesium ions in hydrated cement material chemistry structure based on the mechanical stiffness properties of the modified material chemistry configuration using computational material modeling and simulations. Different computational simulation methods have been employed in the past to study heterogeneous materials. Manzano et al\cite{6} used a combination of forcefield and DFT atomistic simulations, while others have used ab initio methods to study the structure of C-S-H\cite{8,9}. There are also prior studies on the mechanical behavior of heterogeneous materials using multi-scale modeling, and studies to understand cement hydration with finite element modeling, as well as temperature dependency of the microstructure of cement hydrates\cite{3,10,11} reported in the literature. The present paper presents and discusses results from our work on material chemistry modeling to understand the effect of Calcium ion substitution with Magnesium on chemistry level structure of C-S-H on the predicted mechanical stiffness.

2 MATERIAL CHEMISTRY LEVEL MODELING OF C-S-H

Material chemistry level modeling following the principles and techniques commonly grouped under Computational Material Science is one of the key facilitator of material science that is experiencing fast growth pace and forms the basis of the present work\cite{13}. The major hydrated component of cement paste C-S-H gel is responsible for the load bearing attributes and strength of cement, and has a complex material chemistry\cite{14}. The molecular structural representations of C-S-H gel are still inconclusive with several structural representations cited in literature\cite{15}. These include:

- Wollastonite group which comprises of Foshagite (Ca$_4$(Si$_3$O$_9$)(OH)$_2$)$_2$\cite{16}, Hillebrandite (Ca$_2$(SiO$_3$)(OH)$_2$)$_2$\cite{17}, Xonotlite (Ca$_6$Si$_6$O$_{17}$(OH)$_2$)$_2$\cite{18}, Okenite ([Ca$_8$(Si$_6$O$_{16}$)(Si$_6$O$_{15}$)$_2$(H$_2$O)$_6$]$^+$[Ca$_2$(H$_2$O)$_6$]3H$_2$O)]\cite{19}, and others that are compiled by
Richardson[15].

- Tobermorite group which comprises of Clinotbermorite\(^e\) (\(\text{Ca}_5\text{Si}_6\text{O}_{17} \cdot \text{5H}_2\text{O}\)), Clinotbermorite\(^d\) (\(\text{Ca}_5\text{Si}_6\text{O}_{17} \cdot \text{5H}_2\text{O}\)), Clinotbermorite 9Å\(^c\) (\(\text{Ca}_4\text{Si}_6\text{O}_{16} (\text{OH})_2\)), Clinotbermorite 9 Å\(^d\) (\(\text{Ca}_4\text{Si}_6\text{O}_{16} (\text{OH})_2\))[20,21], anomalous and normal Tobermorite 11Å\(^o\) (\(\text{Ca}_4\text{Si}_6\text{O}_{15} (\text{OH})_2 \cdot \text{5H}_2\text{O}\) and \(\text{Ca}_4\cdot\text{5Si}_6\text{O}_{16} (\text{OH})_2 \cdot \text{5H}_2\text{O}\) respectively)[22,23] and Tobermorite 14Å[24].

- Jennite group comprising Jennite and Metajennite (\(\text{Ca}_9\text{Si}_6\text{O}_{18}(\text{OH})_6 \cdot \text{8H}_2\text{O}\)). A complete composition and crystal information for most of the C-S-H and other related phases have been tabulated by Richardson[15].

From various material chemistry molecular structures of C-S-H, Tobermorite 14Å and Jennite structures are traditionally accepted, widely used, and adapted material molecular structural representation of C-S-H cement paste[2,14,15,25,26]. In the present paper, the structural configuration of C-S-H Jennite is employed to study and understand the influence of material chemistry changes due to Magnesium replacement.

Jennite is a representative mineral form of calcium silicate hydrate (C-S-H). It has a chemical formula of \(\text{Ca}_9\text{Si}_6\text{O}_{18}(\text{OH})_6 \cdot \text{8H}_2\text{O}\)[15,24,25]. The molecular/chemistry level structure of C-S-H Jennite is defined by a triclinic unit cell of dimensions \(a = 10.6 \text{ Å}, b = 7.3 \text{ Å}, c = 10.9 \text{ Å}\), and angles \(\alpha = 101.3^\circ, \beta = 97.0^\circ, \gamma = 109.7^\circ\). In Figure 2 a crystal of C-S-H Jennite, consisting of 64 unit cells in a \(4 \times 4 \times 4\) arrangement, is illustrated. The molecular structure of C-S-H Jennite is layered, formed by sheets of calcium oxide connected to short silica chains, three silica monomers\(^1\) long, that are not connected to each other. The calcium oxide sheets are linked to each other by additional calcium octahedral sites. The layers containing the calcium octahedral sites connecting the calcium oxide sheets also contain most of the free water molecules in the structure[25,27].

The atomic positions of the initial unit cell of C-S-H Jennite were obtained from American Mineralogist Crystal Structure Database (AMCSD), which corresponds to the C-S-H Jennite structure defined by Bonaccorsi et al.[25]. The molecular dynamics modeling analysis was performed using the molecular dynamics analysis code Accelrys Material Studio\(^c\), employing the Discover module with the associated atom interaction energy represented by COMPASS force field[28].

Material chemistry molecular models of C-S-H are employed based on a Molecular Dynamics (MD) modeling methodology to understand the influence of Magnesium ion exchange are based on a 64 unit cell configuration shown in Figure 2. The material molecular chemistry system size of 64 unit cells and other MD modeling parameters were determined after a careful study[29]. Other MD simulation analysis parameters employed were a dynamic analysis time of 200 ps; periodic boundary conditions applied to the molecular system, temperature and pressure control based on Nose Thermostat and Parinello Barostat, with a time integration time step of 1 fs (femto second) following a Velocity-Verlet method.

---

\(^{1}\) A silica monomer is a single \(\text{SiO}_2\) unit.
Magnesium Ion Exchange in C-S-H Jennite

The relevance and applicability of Magnesium ion exchange in hydrated cement C-S-H structure was discussed earlier. Following this, Magnesium (Mg) ion exchange in traditional C-S-H Jennite was obtained by replacing the Calcium (Ca) ions in the molecular structure of Jennite C-S-H structure with magnesium ions. The unit cell configuration of C-S-H Jennite consists of 9 calcium ions. The magnesium modified C-S-H Jennite structure based on a same crystal configuration is obtained by replacing one or more calcium atoms by magnesium atoms. The different magnesium modified C-S-H Jennite structures, and their corresponding magnesium percentage based on calcium and magnesium molar mass considered in the present work are tabulated in Table 1.

<table>
<thead>
<tr>
<th>Structure</th>
<th>Ca atoms</th>
<th>Mg atoms</th>
<th>Magnesium percentage by weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>J0</td>
<td>9</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>J1</td>
<td>8</td>
<td>1</td>
<td>7.0</td>
</tr>
<tr>
<td>J2</td>
<td>7</td>
<td>2</td>
<td>14.8</td>
</tr>
<tr>
<td>J3</td>
<td>6</td>
<td>3</td>
<td>23.0</td>
</tr>
<tr>
<td>J4</td>
<td>5</td>
<td>4</td>
<td>32.7</td>
</tr>
<tr>
<td>J5</td>
<td>4</td>
<td>5</td>
<td>43.0</td>
</tr>
<tr>
<td>J6</td>
<td>3</td>
<td>6</td>
<td>54.8</td>
</tr>
<tr>
<td>J7</td>
<td>2</td>
<td>7</td>
<td>68.0</td>
</tr>
<tr>
<td>J8</td>
<td>1</td>
<td>8</td>
<td>82.9</td>
</tr>
<tr>
<td>J9</td>
<td>0</td>
<td>9</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Different unit cell configurations of traditional and magnesium modified C-S-H Jennite as noted above was replicated to create a 64 unit cell configuration in a 4×4×4 arrangement. Figure 3 shows the material chemistry molecular structures of 64 unit cell C-S-H structures created and analyzed in the present work. MD dynamic analysis was performed to obtain...
associated Virial stresses, potential energy of the system and elastic stiffness matrix. The time average values of the molecular dynamic analysis after reaching a stable energy configuration were employed in determining these physical parameters.

### 2.2 Predicted Mechanical Stiffness Properties from Material Chemistry Structure

Mechanical stiffness properties (Bulk, Shear, and Young’s Modulus) were determined from the elastic constants obtained from the second derivative of potential energy of the material chemistry molecular system. This is defined by equation 1.

$$C_{ij} = \frac{1}{V} \frac{\partial^2 U}{\partial \varepsilon_i \partial \varepsilon_j}$$  \hspace{1cm} (1)

Here, $\varepsilon_i$, $\varepsilon_j$ are lattice strain components, $U$ is potential energy, $V$ is simulation cell volume. The $6 \times 6$ elastic stiffness matrix, $C$ can be obtained from the above relation.

$$C_{ij} = \begin{pmatrix} C_{11} & \cdots & C_{16} \\ \vdots & \ddots & \vdots \\ C_{61} & \cdots & C_{66} \end{pmatrix}$$  \hspace{1cm} (2)

Elastic compliance matrix, $S = C^{-1}$ can also be obtained.

$$S_{ij} = \begin{pmatrix} S_{11} & \cdots & S_{16} \\ \vdots & \ddots & \vdots \\ S_{61} & \cdots & S_{66} \end{pmatrix}$$  \hspace{1cm} (3)

Hence the mechanical stiffness properties for the homogeneous stiffness modulus can be predicted following Voight, Reuss or Hill average elastic stiffness relations from the elastic and compliance matrix coefficients as $[30]$:

$$K_R = \frac{1}{(S_{11}+S_{22}+S_{33})+2(S_{12}+S_{23}+S_{31})}$$  \hspace{1cm} (4)

$$K_v = \frac{(C_{11}+C_{22}+C_{33})+2(C_{12}+C_{23}+C_{31})}{9}$$  \hspace{1cm} (5)
\[ K_H = \frac{(K_V + K_R)}{2} \]  
\[ G_R = \frac{15}{4(S_{11} + S_{22} + S_{33}) - 4(S_{12} + S_{23} + S_{31}) + 3(S_{44} + S_{55} + S_{66})} \]  
\[ G_V = \frac{15}{2(C_{11} + C_{22} + C_{33}) - (C_{12} + C_{23} + C_{31}) + 3(C_{44} + C_{55} + C_{66})} \]  
\[ G_H = \frac{(G_V + G_R)}{2} \]  
\[ E = 2G(1 + v) \]

where; \( K_R \) = Reuss bulk modulus, \( K_V \) = Voight bulk modulus, \( K_H \) = Hill average bulk modulus, \( G_R \) = Reuss shear modulus, \( G_V \) = Voight shear modulus, \( G_H \) = Hill average shear modulus, \( E \) = Young’s Modulus, and \( C_{ij} \) and \( S_{ij} \) are the components of the stiffness and compliance matrix respectively \[^{30,31}\].

3 RESULTS AND DISCUSSIONS

The MD analysis methodology and predictive mechanical stiffness values from the potential energy as defined in section 2 were obtained for traditional and different Mg-modified C-S-H material chemistry configurations listed in Table 1. Figure 4 shows the change in the total energy for traditional and Mg-modified C-S-H Jennite configurations. The variation of normalized total energy is shown in Figure 4.

The predicted elastic modulus (Hill average) values for traditional and Mg-modified C-S-H Jennite material chemistry structures are shown in Figure 5. A sharp increase in the predicted elastic modulus is observed for the Mg-modified C-S-H Jennite structure at 33% Mg structure followed by a reduction in predicted elastic modulus values as the percentage of Mg modification further increases. Clearly, these results show that predicted elastic modulus have an impact with variations as shown in Figure 5. A Magnesium ion percentage of 33% is noticed to provide the highest increase in the predicted elastic modulus in Mg-Ion modified C-S-H compared to traditional C-S-H material chemistry structure.

![Figure 4: Change in total energy per mole for traditional and Mg-modified C-S-H Jennite](image-url)
The corresponding variation in the predicted bulk and shear modulus for various Mg-modified C-S-H Jennite material chemistry structures is shown in Figure 6. Again in both cases, present results indicate that there is a positive impact with increased bulk and shear modulus values until a Magnesium ion exchange of about 33%.

4 CONCLUSIONS

Material chemistry and modifications influence the evolving behavior of complex, hierarchical materials such as cement where material chemistry changes influence the hydration process, altering the microstructure and finally the engineering scale properties. In this paper, mechanical stiffness properties based on elastic modulus predictively obtained from traditional and Mg-modified showed that material chemistry does influence structural material stiffness modulus properties, and varies as the amount of Mg inclusion increases. These predicted material properties are based on molecular dynamics based material modeling with material chemistry structures; and thus provides an effective computational methodology to understand the expected variations due to material chemistry changes such as due to Magnesium ion exchange of C-S-H Jennite discussed in the present paper. Results from our current analysis is presented and further investigations are needed to explain some of the non-physical computational model values noted in the predicted values of bulk modulus that are
based on mechanics based approximations following the compliance matrix definitions in MD simulation analysis.

Engineering scale properties of interest are the stress-strain deformation behavior and effective prediction of estimated stress-strain behavior due to material chemistry changes, such as those discussed due to magnesium exchange in C-S-H Jennite are also of interest. Our work and results from such material chemistry level modeling will be highlighted in the presentation and will be discussed in future publications.

REFERENCES

A MULTI-SCALE ELECTRO-CHEMO-MECHANICAL ANALYSIS OF SOLID OXIDE FUEL CELL CONSIDERING EVOLUTION OF MICROSTRUCTURE IN POROUS ELECTRODE

M. MURAMATSU ¹, S. TAKASE ², K. YASHIRO ¹, T. KAWADA ¹ AND K. TERADA ³

¹ Graduate School of Environmental Studies, Tohoku University
468-1-404, Aoba, Aramaki, Aoba-ku, Sendai 980-0845 Japan
e-mail: muramatsu@ee.mech.tohoku.ac.jp

² Department of Civil and Environmental Engineering, Tohoku University
468-1-404, Aoba, Aramaki, Aoba-ku, Sendai 980-0845 Japan
e-mail: takase@irides.tohoku.ac.jp

³ International Research Institute of Disaster Science, Tohoku University
468-1-404, Aoba, Aramaki, Aoba-ku, Sendai 980-0845 Japan
e-mail: tei@irides.tohoku.ac.jp

Key words: Solid Oxide Fuel Cells, Potential Simulation, Deformation Analysis, Microstructure.

Abstract. In the multi-scale analysis of oxygen potential distributions in Solid Oxide Fuel Cell (SOFC), the evolution of the microstructure in anode is reflected in the macroscopic electrical conductivities and the amount of triple-phase boundaries which are the generation sites of electro-chemical reaction currents. The oxygen potential distributions induce the reduction strains in oxide materials along with the thermal strains due to temperature change. These strains cause the macroscopic deformation in mutually constrained components. The capability of the proposed method is demonstrated in characterizing the aging degradation of the macroscopic electro-chemo-mechanical behaviour of SOFC that is caused by the Ni-sintering in anode microstructures during long-period control.

1 INTRODUCTION

Components of Solid Oxide Fuel Cells (SOFCs) are always exposed to high temperature and large gas pressure under operation. Subjected to the starting and stopping control, the performance of SOFC is gradually degraded. One of the degradation factors is the thermal expansive deformation due to temperature change [1]. Another factor is the expansive deformation under reduction environment, which is characteristic deformation for SOFC [1]. These stress-free deformations inevitably invoke the unexpectedly deformation due to the mutual constraints of the components, which sometimes causes the mechanical deterioration. Another notable point about degradation in SOFC is the aggregation of Ni particles in the porous anode electrode of SOFC under high temperature environment. The temporal changes of the anode microstructures cause the degradation of not only the overall mechanical properties, but also the electrical performance.
In order to realize the macroscopic electro-chemo-mechanical coupling analyses of an SOFC under operation, we propose a characterization method of the time-varying overall or macroscopic electro-chemical and mechanical properties of anode electrodes by employing the phase-field method that enables us to capture the time-varying geometry of anode microstructures due to Ni-sintering. On the basis of the information of the microstructures, the homogenization method is adopted to evaluate the temporal change of the macroscopic electro-chemo-mechanical properties that characterize the macroscopic inelastic mechanical behavior and the oxygen potential distribution in SOFC.

The framework of the proposed method in this study is shown in Fig. 1. For the chemical-electro simulation of oxygen potential distributions, the time-evolution of the anode microstructure is reflected in not only the macroscopic electrical conductivities, but also the amount of triple-phase boundaries, which are the generation sites of electro-chemical reaction currents. Once the oxygen potential distributions are determined, the time-variation of the reduction-induced strains in oxide materials can be calculated. The thermal strains due to temperature also change according to the change in microstructure. Both of these strains cause the macroscopic deformation in mutually constrained components. In the mechanical field, elastic and creep strains are calculated considering microstructure of anode in SOFC. The numerical material testing is used for the homogenization of microscopic quantities [2]. The proposed method is demonstrated in characterizing the aging degradation of the macroscopic electro-chemo-mechanical behavior of SOFC that is caused by the Ni-sintering in anode microstructures during long-period control.

2 FORMULATION

2.1 Electro-chemical phenomenon

At the cathode electrode, oxygen $O_2$ is reduced by electron $e^-$ to produce oxygen ion $O^{2-}$,
which can move through the electrolyte toward the anode electrode. The oxygen ion that reaches the anode electrode is oxidized by the reaction with H₂, which is provided as a fuel, to produce water H₂O and e⁻.

\[
\text{Cathode: } \frac{1}{2} \text{O}_2 + 2 \text{e}^- \rightarrow \text{O}^{2-} \quad (1)
\]

\[
\text{Anode: } \text{O}^{2-} + \text{H}_2 = \text{H}_2\text{O} + 2 \text{e}^- \quad (2)
\]

Since the electrodes are porous materials, actual reactions are expected to occur at the interfaces between gas and solid phases [1]. The latter electro-chemical reaction, i.e., oxidation with emission of electrons, occurs at the so-called triple phase boundaries (TPBs) of the anode electrode, which is generally a cermet composed of Ni particles and ion conductive ceramics. On the other hand, the cathode electrode, whose constituents are mixed ionic-electronic conductive ceramics, allows the reduction of O₂ at the so-called dual phase boundaries (DPBs), involving the ionization of the oxygen with absorption of electrons.

The electronic current density \( i_{\text{Anode}} \) generating with the oxidation (2) is known to be determined by the Butler-Volmer relationship [3,4] as

\[
i_{\text{Anode}} = i_0^{\text{Anode}} \exp \left\{ \frac{(2\alpha_a F)}{(R\theta)\Delta E} - \exp \left\{ \frac{(2\alpha_c F)}{(R\theta)\Delta E} \right\} \right\} \quad (3)
\]

where \( F = 96500 \text{Cmol}^{-1} \) is the Faraday’s constant, \( R \) is the gas constant, \( \theta \) is the temperature and \( i_0^{\text{Anode}} \) is the exchange current density, \( \alpha_a \) and \( \alpha_c \) are the anodic and cathodic charge transfer coefficients. In addition, \( \Delta E \) is the polarization voltage expressed by

\[
\Delta E = \left( \mu_0 - \mu_0^{\text{gas}} \right) / 2F \quad (4)
\]

where \( \mu_0 \) and \( \mu_0^{\text{gas}} \) are the oxygen potentials in material and in gas, respectively. The oxygen potential is calculated with the partial pressure of oxygen \( p_{\text{O}_2} \) as

\[
\mu_0 = \mu_0^0 + \frac{1}{2} R\theta \ln p_{\text{O}_2} \quad (5)
\]

where \( \mu_0^0 = 0 \) is the oxygen potential at standard condition. The normal component of current density \( i \) inside the Ni particles equals to the generated electric current density as

\[
i = i_{\text{Anode}} \cdot n = i_{\text{Anode}} \quad (6)
\]

where \( n \) is the outward unit vector on the TPB, directed to the normal to Ni surface. According to the electro-chemical reaction Eq. (2), the oxygen ionic current density generated at the same boundary surfaces can be evaluated as

\[
i_{\text{O}^{2-}} = i_{\text{O}^{2-}} \cdot n' = i_0^{\text{Anode}} = -i_{\text{Anode}} \quad (7)
\]

where the outward unit vector \( n' \) is defined to the TPB, directed to the normal to the oxide’s surface.

The current density which generates at the cathode is also determined by the Butler-Volmer type equation as

\[
i_{\text{Cathode}} = i_0^{\text{Cathode}} \exp \left\{ \frac{(2(1 - \alpha) F)}{(R\theta)\Delta E} - \exp \left\{ \frac{(2\alpha F)}{(R\theta)\Delta E} \right\} \right\} \quad (8)
\]

where the exchange current density \( i_0^{\text{Cathode}} \) and the charge transfer coefficient \( \alpha \) are different
from those for anode. The relationships of this generated current density with the normal components of the current densities in the oxide are given as

\[ i_e = i_{e}\cdot n = i_{e}\text{Cathode} \] (9)

\[ i_{O^2^-} = i_{O^2^-}\cdot n' = i_{O^2^-}\text{Cathode} = -i_{e}\text{Cathode} \] (10)

### 2.2 Microscopic electronic and oxygen-ionic conduction problems

The electrical conduction in mixed ionic-electronic conductive ceramics can be characterized by the local transport of electro-chemical potentials and is governed by the following equations [5]:

\[ \frac{e}{2F} \mu_{O^2^-} = -\nabla \cdot i_{O^2^-} - \frac{\sigma_{O^2^-}}{2F} \nabla \eta_{O^2^-} \] (11)

\[ \frac{e}{2F} (-\mu_{O^2^-}) = -\nabla \cdot i_e - \frac{\sigma_e}{2F} (\nabla 2 \eta_e) \] (12)

\[ \mu_{O^2^-} = 2\eta_{e} - i = i_{O^2^-} + i_e \] (13)

where \( \eta_{O^2^-} \) is the electro-chemical potential of oxygen ion, \( \eta_e \) is the electro-chemical potential of electron, \( c = c(\mu_{O^2^-}) = (-4F^2/V_m)(\partial \delta(\mu_{O^2^-})/\partial \mu_{O^2^-}) \) is the electric capacitance, \( \delta \) is the oxygen vacancy, \( V_m \) is the molal volume, \( i \) is the total current density, and \( \sigma_{O^2^-} \) and \( \sigma_e \) are the electric conductivities of oxygen ion and electron, respectively. It is noted that the source terms of reaction current densities are not appeared in Eqs. (11) and (12). The reaction currents are generated at the TPBs or DPBs in the macroscopic scale. The standard Dirichlet boundary conditions with the values of electro-chemical potentials are used for the pore surface in the microscopic problem.

<table>
<thead>
<tr>
<th>Table 1 Electric conductivities and oxygen vacancy</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sigma_e )</td>
</tr>
<tr>
<td>LSCF (Cathode)</td>
</tr>
<tr>
<td>8YSZ (Electrolyte)</td>
</tr>
<tr>
<td>Ni-YSZ (Anode)</td>
</tr>
<tr>
<td>LCCr (Interconnect)</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
2.3 Macroscopic electronic and oxygen-ionic conduction problems

The local transport of macroscopic electro-chemical potentials is governed by the following equations:

\[
\begin{align*}
- \frac{\varepsilon}{2F} \hat{\mu}_o &= -\nabla \cdot \hat{i}_{o^-} + \hat{j}_{o^-}, \quad \hat{i}_{o^-} = -\frac{\sigma_{o^-}}{2F} \nabla \eta_{o^-} \\
- \frac{\varepsilon}{2F} (-\hat{\mu}_o) &= -\nabla \cdot \hat{i}_e + \hat{j}_e, \quad \hat{i}_e = -\frac{\sigma_e}{2F} (\nabla 2 \eta_e) \\
\hat{\mu}_o &= \hat{\eta}_{o^-} - 2 \hat{\eta}_e, \quad \hat{i} = \hat{i}_{o^-} + \hat{i}_e
\end{align*}
\]

(14)  
(15)  
(16)

where symbol \(\bullet\) indicates the macroscopic (homogenized) quantity. The reaction current densities \(\hat{j}_{o^-}\) and \(\hat{j}_e\) are appeared as source terms in Eqs. (14) and (15). Once the solutions of these equations \(\eta_{o^-}\) and \(\eta_e\) are determined, the macroscopic oxygen potential \(\hat{\mu}_o\) can be obtained, then the macroscopic vacancy \(\delta\) is evaluated by the multiplication of from Table 1.

2.4 Macroscopic deformation problem

The macroscopic (homogenized) governing equations of deformation problem are written in the following forms:

\[
\nabla \cdot \mathbf{T} + \rho \mathbf{b} = 0
\]

(17)

\[
\mathbf{\ddot{e}} = \mathbf{\ddot{e}}^e + \mathbf{\ddot{e}}^e + \mathbf{\ddot{e}}^i + \mathbf{\ddot{e}}^0
\]

(18)

\[
\mathbf{T} = \mathbf{C} : \mathbf{\dot{e}}^e = \kappa (\text{tr} \mathbf{\dot{e}}^e) \mathbf{I} + 2 \mu \mathbf{\dot{e}}^e
\]

(19)

\[
\mathbf{\ddot{e}}^e = \mathbf{\ddot{e}}^T_m / \||\mathbf{T}_m||, \quad \mathbf{\ddot{e}}^i = \mathbf{\ddot{e}}^i / \||\mathbf{T}||^2 \exp(-C_j / \tilde{\sigma})
\]

(20)

\[
\mathbf{\ddot{e}}^0 = \mathbf{\ddot{a}} \Delta \mathbf{\theta}
\]

(21)

\[
\mathbf{\ddot{e}}^i = \mathbf{\ddot{b}} \Delta \mathbf{\ddot{\delta}}
\]

(22)

where symbol \(\bullet\) indicates the macroscopic (homogenized) quantity. The macroscopic reduction-induced strain \(\varepsilon^i\) is evaluated by the multiplication of the macroscopic vacancy \(\delta\) by the coefficient \(\beta\). The standard Dirichlet and Neumann boundary conditions are used in the deformation problem.

2.5 Microstructures of anode

In this study, we assume that the main factor that deteriorates the anode made by Ni-YSZ is the sintering of Ni particles, which aggregate during the steady operation at 973-1173K. In particles touching with each other at lower temperature than the melting point, material transfer occurs so as to reduce the surface energy of the system. With the driving force of the surface energy, the shapes of particles change with time. To simulate such a time-variation, a phase-field model for sintering [6,7] is employed in this study. We employ the following Ginzburg-Landau type free energy [8]:

\[
F = \int \left[ \sum_i \frac{1}{2} \alpha_i ||\nabla \phi_i ||^2 \right]
\]
\[ + \sum_i A_i \phi_i^2 (1 - \phi_i)^2 + \sum_i B_i \phi_i^2 + \sum_i \left( \sum_{j=1}^3 \frac{\beta_{ij}}{4} \phi_i^2 \phi_j^2 \right) + \frac{\gamma}{2} \phi_i^2 \phi_j^2 \right] dV \]  \hspace{1cm} (23)

where index \( i \) takes 1, 2 or 3, which indicates the phase of Ni, YSZ and void. Here, the first and second terms are the surface energies, the third term is the chemical free energy, and the fourth term is the interface energy of the TPB. Also, coefficients \( \alpha_i, A_i, B_i, \beta_{ij} = \beta_{ji} \) and \( \gamma \) are the material constants associated with the corresponding energies or potentials. The Cahn-Hilliard equations are used for the evolution equation because \( \phi_i \) is a conservative quantity.

\[
\frac{\partial \phi_i}{\partial t} = \nabla \cdot \left( M_i(\phi_i, \theta) \left( \nabla \frac{\delta F}{\delta \phi_i} \right) \right) \]  \hspace{1cm} (24)

where \( M_i(\phi_i, \theta) \) is the mobility defined as \( M_i(\phi_i, \theta) = D(\phi_i) m_i(\theta) \) with the diffusivity function \( D(\phi_i) \) and the mobility function \( m_i(\theta) \), i.e., \( D(\phi_i) = D_{\text{vol}} f(\phi_i) + D_{\text{vap}} (1 - f(\phi_i)) + D_{\text{surf}} \phi_i (1 - \phi_i) \) and \( m_i(\theta) = 10^8 / [1 + \tanh(\theta - \alpha_i / b_i)] \) where \( D_{\text{vol}}, D_{\text{vap}} \) and \( D_{\text{surf}} \) are the diffusivities for the volumetric diffusion, the gas diffusion and the surface diffusion, respectively, and \( \alpha_i \) and \( b_i \) are the mobility parameters and \( f(\phi_i) \) takes the form of \( f(\phi_i) = \phi_i^4 (10 - 15\phi_i + 6\phi_i^5) \). Substitution of \( M_i(\phi_i, \theta) \) into Eq. (24) yields the following equation:

\[
\frac{\partial \phi_i}{\partial t} = m_i(\theta) \left( \nabla D_i(\phi_i) \cdot \left( \nabla \frac{\delta F}{\delta \phi_i} \right) + D_i(\phi_i) \left( \nabla^2 \frac{\delta F}{\delta \phi_i} \right) \right) \]  \hspace{1cm} (25)

Discretizing Eq. (25) with finite differential method, we simulate the development of order parameters.

3 ANALYSIS CONDITION

3.1 Material constants

We employ \( \text{La}_{0.6}\text{Sr}_{0.4}\text{Co}_{0.2}\text{Fe}_{0.8}\text{O}_3-\delta \) (LSCF) for the cathode, 8mol% yttria-stabilized zirconia (8YSZ) for the electrolyte, Ni-YSZ cermet for the anode \( \text{La(\text{Ca})CrO}_3 \) (LCCr) for the interconnect in this study. Note that the formulation does not limit the materials of cells. To evaluate the macroscopic material properties of the porous anode and cathode electrodes by performing the homogenization analyses, homogenization method are used to obtain the macroscopic material properties. In this study, numerical material testing [2] is employed for homogenization. Especially, the microstructure of anode, which is Ni-YSZ cermet, changes temporally due to Ni-sintering. The macroscopic properties of anode are calculated on the basis of the temporal microstructures.

3.2 Temporal change of microstructure in anode

Figure 2 shows the microscopic analysis models of anode (Ni-YSZ) and cathode (LSCF) microstructures, which are generated by the 3-dimensional porous material simulator POCO [9], though other methods such as [10] are available. A phase-field simulation with the above-described evolution equations is carried out only for the Ni particles in Fig. 3 (a) that are assumed to be exclusively sintered. The finite difference model is generated with 5050 146 grids \((\Delta x \times \Delta y \times \Delta z = 0.062\mu m \times 0.062\mu m \times 0.062\mu m)\). The material constants are chosen as \( \alpha_1 = 14.0 \times 10^2 \text{ J/cm}^3, A_1 = A_2 = 22.0 \times 10^3 \text{ J/cm}^3, B_1 = B_2 = 1.0 \text{ J/cm}^3, \beta_{12} = 5.0 \times 10^3 \text{ J/cm}^3, \beta \)
\[ \beta_{23} = 0.0 \text{ J/cm}^3, \quad \gamma = 100.0 \times 10^{-3} \text{J/cm}^3, \quad D_{\text{vol}} = 5.0 \times 10^{-13} \text{ cm}^2/\text{s}, \quad D_{\text{vap}} = 5.0 \times 10^{-15} \text{ cm}^2/\text{s} \text{ and } D_{\text{surf}} = 2.0 \times 10^{-10} \text{ cm}^2/\text{s}. \] It is noted here that the diffusivity of the interface is set at a value larger than the ones for the solid and gas phases to promote the sintering near the interfaces.

Figure 2 is the schematics of the condition of temperature control in a virtual SOFC operation. As shown in Fig. 3(a), after the temperature increases to 1073K, this constant value of temperature is kept for 11 hours, and then decreases to the room temperature by natural cooling. Under this temperature control, the sintering simulation is assumed to start when the temperature reaches at 1073K, since Ni particles are sintered mainly in steady operations [1]. The hydrogen gas is supplied for 10 hours as shown in Fig. 3(b). Here, the simulation time is 11h with \( \Delta t = 1 \text{s} \). The results of the sintering simulation of Ni particles are provided in Fig. 4(a) that shows the time-variation of the anode microstructure with the TPBs, which are generation sites of reaction currents. The corresponding time-variation of the amount of the
TPBs is shown in Fig. 4(b), revealing that the frequency of TPBs is reduced with time during the sintering of Ni particles.

### 3.2 Homogenization

With the geometrical and material information about anode, which is shown in section 3.2, and cathode, numerical material tests are conducted to evaluate the macroscopic properties such as oxygen ionic and electronic conductivities, elastic constants, creep parameters and coefficient of thermal expansion. The obtained properties are used for the macroscopic analysis [2].

### 4 ANALYSIS RESULT

#### 4.1 Electric conduction problem

Figure 5 is the analysis model and boundary conditions for macroscopic electric conduction problem. Figures 6 show the time variations of the distributions of the potentials of electrons and oxygen ions, respectively. The contour plot figures show their time varying distributions in the $x_1$-$x_2$ plane. An increase of the uniform distributions of the electrochemical potential of electrons, which can be observed in Figs. 6 (i) and (ii), reflects the supplying condition of fuel in the anode in the numerical analysis. After 3,600s, the potential is kept...
constant as can be seen from Figs. 6 (ii) and (iii), realizing the steady states. Although the effects of Ni sintering have been taken into account in the values of the homogenized mechanical properties, they seem to be little in the present deformation analysis for this cell structure. The effect might be significant depending on the shape design of cells.

4.2 Deformation problem

The analytic model and boundary conditions for macroscopic analysis for the macroscopic deformation problem are shown in Fig. 7. Figure 8 shows the displacement at point P along -x_2 axis. As can be seen from Fig. 8, at first, the overall cell structure is bent toward the interconnect side This initial bending until the temperature reaches at 1073K has been caused by the significant development of larger thermal expansion strains in the cathode than those in the interconnect. Subsequently, the deflection is gradually decreased and the head of the cell moves toward the cathode side during 10h-steady operation and cooling, while the overall shape of the cell is warped into the interconnect side. This is probably caused by the reduction induced expansive strains that are expected to developed in the interconnect side during steady operation and subsequent cooling.

5 CONCLUSIONS

In this study, to perform deformation analyses of SOFCs under operation, a characterization method of its time-varying macroscopic electro-chemo-mechanical behavior of electrodes is developed by taking into account the temporal changes of geometries of anode microstructures due to Ni-sintering. The conclusions are summarized as follows:

- A phase-field method was presented to simulate the micro scale morphology change with time. As a result, the time variation of the amount of TPBs was directly predicted. In the homogenization analyses, the dependencies of the properties of constituent materials on temperature and the oxygen potential, which is supposed to change within an operation period, were also considered.

- Under the assumption of an actual start-and-stop operation, the macroscopic electric
conduction analysis was conducted to oxygen potential distribution in an overall cell structure under long-period operation with start-and-stop control, which determined reduction-induced deformation of oxide materials.

- The macroscopic deformation analysis was carried out for the overall SOFC structure with the stress-free strains and the homogenized mechanical properties, both of which depend on the operational environment.

ACKNOWLEDGEMENT

This work was carried out as a part of the research project “Technology development for SOFC commercialization promotion, Basic study on rapid evaluation method of SOFC durability”, which was supported by the New Energy and Industrial Technology Development Organization (NEDO), Japan.

REFERENCES

LOCALIZED VERSUS DIFFUSED FAILURE MODES IN CONCRETE SUBJECTED TO HIGH TEMPERATURE

SONIA M. VRECH\textsuperscript{1}, MARIANELA RIPANI\textsuperscript{2} AND GUILLERMO ETSE\textsuperscript{3}
\textsuperscript{1}CONICET - University of Tucumán
Av. Independencia 1800. 4000 - Tucumán. Argentina
svrech@herrera.unt.edu.ar
\textsuperscript{2}CONICET- University of Buenos Aires
Av. Las Heras 2214. C1127AAR - Buenos Aires. Argentina
mripani@fi.uba.ar
\textsuperscript{3}CONICET - University of Tucumán
Av. Independencia 1800. 4000 - Tucumán. Argentina
getse@herrera.unt.edu.ar

Key words: Discontinuous bifurcation, gradient poroplasticity, temperature

Abstract. In this work, conditions for discontinuous bifurcation in limit states are derived and evaluated in quasi-brittle materials like concrete, for variable stress states, loading conditions and high temperature scenarios. This is performed in the framework of a thermodynamically consistent non-local poroplastic constitutive theory for quasi-brittle materials subjected to high temperature [14]. Thereby, gradient poroplasticity and fracture energy-based homogenization are combined to describe the post-peak behavior of porous media when subjected to long term exposure of temperature. For the numerical implementation of the model, the consistent tangent operator is developed and the dual mixed FE formulation for thermodynamically consistent gradient plasticity by [20] is considered, which was extended to porous media by [10]. The explicit solutions for brittle failure conditions in the form of discontinuous bifurcation are proposed for this poroplastic constitutive theory based on gradient plasticity. The results provide relevant information regarding the variation of the transition point of brittle-ductile failure mode with the acting temperature and confining pressure.

1 INTRODUCTION

Failure behavior of quasi-brittle materials like concrete is characterized by strong spatial discontinuities of the kinematic fields when they are sufficiently deformed into the inelastic regime, depending on the governing stresses and on the mechanical and chemical features of the material micro and mesostructure, see a.o. [23, 18].
Whereas in tensile regime the concrete response is highly brittle as the damage entirely localizes in one single crack of zero width while the material outside the crack remains practically undamaged and subjected to elastic unloading, see a.o. [13, 3], in compressive regime the ductility of concrete failure behavior strongly increases with the confining pressure, governed by both, fracture energy releases in active microcracks and material degradation in between these cracks, see a.o. [8, 17].

This complex variation from brittle to ductile failure modes depending on the acting stress state, is also strongly affected by variable temperature fields. This effect strongly influences the mechanical behavior and the overall strength capacity of the related structures. Regarding the influence of the concrete performance on its sensitivity to high temperatures, many authors agree that high strength concretes reduce their resistance when they are subjected to high temperatures, see a.o. [1, 5].

Based on the first contributions related to the concrete behavior under high temperatures in porous media framework by [16], and in conjunction with the gradient and fracture energy-based Leon - Drucker Prager (LDP) constitutive model for quasi-brittle materials like concrete by [21], the temperature dependent Leon - Drucker Prager model (TD-LDP) have been formulated by the authors, see [14].

In this work, it becomes particularly important the evaluation of brittle or localized failure modes of poroelastic materials that in the framework of continuum models are described in terms of discontinuous bifurcations or jumps in the velocity gradients.

Several analytical and geometrical attempts have been aimed to capture the onset of localization and to determine both direction and amplitude of the related cracks or shear bands. After the original works by [6, 7], many authors studied the problem in a systematic manner, formulating mathematical indicators that signalize the initiation of localized failure modes in the form of discontinuous bifurcation, see a.o. [23, 12]. In the field of quasi-brittle materials can be cited the contributions by [4, 19, 22]. Recently, discontinuous bifurcation analysis of thermodynamically consistent gradient poroelastic materials have been developed by [11].

Analytical methods are applied in this contribution to evaluate the predictions of localized failure modes and the transition from brittle to ductile failure provided by the thermodynamically consistent gradient and fracture energy-based plasticity theory for poroelastic materials like concrete subject to high temperature levels.

2 THERMODYNAMICALLY CONSISTENT GRADIENT BASED POROELASTIC THEORY

The First Thermodynamics Law corresponding to poroelastic gradient materials states that

\[ \dot{E} + \dot{K} = P + Q \]  

(1)
with the internal and kinetic energy rates $\dot{E}$ and $\dot{K}$, the mechanic work of external forces $P$ and the externally supplied heat $Q$ are computed as

$$
\dot{E} = \frac{d}{dt} \int_{\Omega} e \, d\Omega; \quad \dot{K} = \frac{1}{2} \int_{\Omega} [\rho_s (1 - \phi) \dot{u}_s \cdot \dot{u}_s + \rho_f \phi \mathbf{w} \cdot \mathbf{w}] \, d\Omega \tag{2}
$$

$$
P = \int_{\partial\Omega} (\sigma \cdot \mathbf{n}_s \cdot \dot{u}_s - p \rho \mathbf{n}_s \cdot \mathbf{w}) \, d\partial\Omega + \int_{\Omega} \rho \mathbf{b} \cdot \dot{u}_s \, d\Omega \tag{3}
$$

$$
Q = \int_{\Omega} \rho r \, d\Omega - \int_{\partial\Omega} h \cdot \mathbf{n}_s \, d\partial\Omega \tag{4}
$$

being $e = e_s + e_f$ the internal energy density, where the subscripts $s$ and $f$ denote the skeleton and fluid components respectively, $\mathbf{w}$ the fluid mass flux vector, $\rho = \rho_s (1 - \phi) + \rho_f \phi$ the total density of the body and $\phi$, the porosity. Moreover $\mathbf{u}_s$ represents the velocity vector of the skeleton, $\sigma$ the total Cauchy tensor, $\mathbf{b}$ the body forces, $r$ the applied heat sources, $p$ the pore pressure and $\mathbf{h}$ the heat flux vector.

The Second Thermodynamics Law is expressed as

$$
\dot{S} - Q_T \geq 0 \tag{5}
$$

being $\dot{S}$ the system entropy rate and $Q_T$ the entropy flux. Each term in Eq. (5) is defined as

$$
\dot{S} = \int_{\Omega} [\rho \dot{s} + \nabla \cdot (s^m \mathbf{w})] \, d\Omega; \quad Q_T = -\int_{\partial\Omega} \frac{\mathbf{h} \cdot \mathbf{n}_s}{T} \, d\partial\Omega + \int_{\Omega} \frac{\rho r}{T} \, d\Omega \tag{6}
$$

being $s$ the internal entropy density, $s^m$ the fluid internal entropy per unit of mass and $T$ the absolute temperature. According to [2], the weak form of the Second Thermodynamics Law can be obtained as

$$\int_{\Omega} \left[ \sigma : \dot{\varepsilon} + g^m \dot{m} + \rho T \dot{s} - \rho \dot{\varepsilon} + \mathbf{w} \cdot (\nabla s^m - \nabla h_f) - \frac{\mathbf{h}}{T} \cdot \nabla T \right] \, d\Omega \geq 0 \tag{7}
$$

where $g^m$ is the enthalpy density per unit mass, $h_f = e^m_f + p/\rho_f$ corresponds to the fluid specific enthalpy, $e^m_f$ is the fluid internal energy per unit of mass and $\varepsilon$ is the strain tensor.

Regarding Eq. (7) and taking into account that an arbitrary thermodynamic states of an open gradient poroplastic material under non-isothermal condition may be defined in terms of the elastic strain tensor $\varepsilon^e = \varepsilon - \varepsilon^p$, the elastic fluid mass content $m^e = m - m^p$, the elastic entropy $s^e = s - s^p$ and the scalar internal variable $q$, in case of isotropic plasticity, the Coleman’s equations result

$$
\sigma = \rho \frac{\partial \psi_s}{\partial \varepsilon^e}, \quad p = \rho \frac{\partial \psi_s}{\partial m^e}, \quad s_s = -\frac{\partial \psi_s}{\partial T}, \quad Q^f = -\rho \frac{\partial \psi_s}{\partial q}, \quad Q^{nl} = T \nabla \cdot \left( \frac{\rho}{T} \frac{\partial \psi_s}{\partial q} \right) \quad \text{in the domain } \Omega \tag{8}
$$
Whereas the dissipations related to the plastic process, heat and fluid mass transports can be computed as
\[
\varphi^p = \sigma : \dot{\varepsilon}^p + \frac{p}{\rho_f} \dot{m}^p + Q^l \geq 0; \quad \varphi^{th} = -\frac{h \cdot \nabla T}{T} \geq 0; \quad \varphi^f = w \cdot (\nabla s^m - \nabla h_f) \geq 0 \quad (10)
\]
To achieve the above equations from Eq. (7), it is assumed that the total Helmholtz’s free energy is decomposed as
\[
\psi = e - Ts = \psi_s + m \psi_f \quad \text{and the total entropy is decomposed}
\]
\[
s = s_s + m s_f, \quad \text{where } \psi_s \text{ and } \psi_f \text{ are the solid- and the fluid-specific Helmholtz’s free energy}
\]
respectively, \( s_s \) and \( s_f \) are the specific entropies related to the solid skeleton and the fluid part respectively, and \( Q = Q^l + Q^{nl} \) is the total dissipative stresses, with \( Q^l \) the local dissipative stress and \( Q^{nl} \) the non-local gradient one.

Adopting the following additive expression for the Helmholz’s free energy
\[
\psi_s (\varepsilon^e, m^e, T, q, \nabla q) = \psi^e (\varepsilon^e, m^e, T) + \psi^f (q, T) + \psi^{nl} (\nabla q), \quad (11)
\]
the elastic, local and non-local plastic counterparts of the free energy are given by
\[
\rho \psi^e = \frac{1}{2} \varepsilon^e : E : \varepsilon^e + \frac{1}{2} M \left( \frac{m^e}{\rho_f} \right)^2 - \frac{1}{2} \chi T^2 + \kappa m^e T - \frac{m^e}{\rho_f} \cdot MB : \varepsilon^e - TA : \varepsilon^e \quad (12)
\]
\[
\rho \psi^l = \frac{H}{2} q^2 - Ts_{fr}^{(q)}; \quad \rho \psi^{nl} = \frac{1}{2} l_c^2 \nabla q \cdot H^{nl} \cdot \nabla q \quad (13)
\]
being \( E \) the fourth order elastic tensor, \( M \) the Biot’s modulus, \( \chi \) the concrete heat capacity, \( \kappa \) the latent heat of variation in fluid mass content, \( B = bI \) the Biot’s tensor with \( b \) the Biot’s coefficient and \( I \) the second order identity tensor, and \( A = \alpha T I \) the thermal expansion tensor, with \( \alpha T \) the thermal expansion coefficient.

In Eq. (13-a), \( s_{fr}^{(q)} \) represents the unrecovered entropy called frozen entropy, see [2]. Moreover, in Eq. (13-b) \( H \) is the local-plastic hardening/softening modulus, \( H^{nl} \) the gradient softening second-order tensor and \( l_c \) the gradient characteristic length. In the particular case of gradient isotropy, the gradient softening second-order tensor can be expressed as \( H^{nl} = H^{nl} I \), being \( H^{nl} \) a positive non-zero scalar.

### 2.1 Constitutive equations

Summarizing, the constitutive equations in terms of total stresses, pore pressure, fluid-specific internal entropy and dissipative stresses for thermodynamically consistent gradient poroplastic materials subjected to non-isothermal conditions are
\[
\sigma = E : \varepsilon^e - M \frac{m^e}{\rho_f} B - T A \quad (14)
\]
\[
p = \frac{M}{\rho_f} m^e - M B : \varepsilon^e + \rho_f \kappa T; \quad s_s = A : \varepsilon^e - \kappa m^e + \chi T + s_{fr}^{(q)} \quad (15)
\]
\[
Q^l = -Hq + T \frac{\partial s_{fr}^{(q)}}{\partial q}; \quad Q^{nl} = l_c^2 H^{nl} \left( -\frac{\nabla T}{2T} \cdot \nabla q + \nabla^2 q \right) \quad (16)
\]
2.2 Non-local poroplastic flow rule

For general non-associated flow rule and hardening/softening laws, the yield surface \( \Phi \) and the dissipative potential \( \Phi^* \) are introduced. The rate of the plastic strains tensor, plastic mass and internal variables are defined as

\[
\dot{\varepsilon}^p = \dot{\lambda} \frac{\partial \Phi^*}{\partial \sigma}, \quad \dot{m}^p = \dot{\lambda} \frac{\partial \Phi^*}{\partial p}, \quad \dot{q} = \dot{\lambda} \frac{\partial \Phi^*}{\partial Q}
\]

being \( \dot{\lambda} \) the rate of the plastic multiplier. To complete the problem formulation, the Kuhn-Tucker conditions and poroplastic consistency must be introduced (See a detailed explanation in [14])

\[
\dot{\lambda} \geq 0, \quad \Phi(\sigma, p, Q) \leq 0, \quad \dot{\lambda} \Phi(\sigma, p, Q) = 0; \quad \dot{\Phi} = 0
\]

3 SOLUTIONS FOR LOCALIZED FAILURE OF POROUS MEDIA

In the realm of the smeared crack approach, localized failure modes are related to discontinuous bifurcations of the equilibrium path, and lead to lost of ellipticity of the equations that govern the static equilibrium problem. Analytical solutions of the discontinuous bifurcation condition in gradient non-local continua has been performed by [15], whereas bifurcation analysis in gradient-based poroplastic media has been carried out by [9], establishing the discontinuous bifurcation condition as either the jump of the velocity gradients

\[
[u_{i,j}] = c_i n_j \quad [\dot{\varepsilon}_{ij}] = \frac{1}{2} (c_i n_j + n_j c_i)
\]

being \( n_i \) the unit vector normal to the characteristic failure surface \( S \), and/or the jump of the rate of fluid mass content

\[
[\dot{m}] = -M_{i,i} = -n_i c_i^M
\]

different to zero. Moreover, the following balance conditions must be considered, see [6, 2]

\[
c[p_{,i}] + [\dot{p}] n_i = 0; \quad c[\sigma_{ij,j}] + [\dot{\sigma}_{ij}] n_j = 0
\]

being \( c \) the discontinuous propagation velocity.

3.1 Bifurcation analysis in drained local porous media

Regarding the drained state in a hardened concrete, taking into account Eq. (21-b) and the evolution of the constitutive equation in Eq. (14), the following condition is obtained

\[
[\dot{\sigma}_{ij}] n_j = Q_{ij} c_j = 0
\]

where \( Q_{ij} \) is the elastoplastic acoustic tensor for local poroplasticity under drained conditions, computed as

\[
Q_{ij} = E^{ep}_{ijkl} n_k n_l = Q^e_{ij} - Q^p_{ij}
\]
with $E^e_{ijkl}$ the solid skeleton elastoplastic tensor, $Q^e_{ij} = E_{ijkl}n_l n_k$ the elastic component of the acoustic tensor, and

$$Q^p_{ij} = \frac{E_{ijkl} f^*_{mn} f_{pq} E_{pqkl}}{h} n_l n_k$$

with $f^*_{mn} = \frac{\partial \Phi^*}{\partial \sigma_{mn}}$, $f_{pq} = \frac{\partial \Phi}{\partial \sigma_{pq}}$ (24)

the plastic one, being $h$ the generalized plastic modulus (see [10] for detailed explanation).

![Figure 1](image)

**Figure 1**: (a) Localization indicator in normal direction to the TD-LDP maximal strength criterion and (b) Transition points for TD-LDP maximal strength criterion.

### 3.2 Bifurcation analysis in drained gradient non-local porous media

Analytical solutions of the discontinuous bifurcation conditions in gradient non-local continua follows from the waves propagation analysis, see [15].

Regarding an homogeneous state before the start of the bifurcation, harmonic perturbations to the incremental field variables, i.e. displacements, mass content, plastic multiplier and temperature are applied, representing the propagation of stationary planar waves

$$\begin{pmatrix}
\dot{U}(t) \\
\dot{\Gamma}(t) \\
\dot{L}(t) \\
\dot{T}(t)
\end{pmatrix} = \exp \left( \frac{i 2 \pi}{\delta} n \cdot x \right)$$

being $\dot{\gamma}$ the mass content rate, $x$ the position vector (in Cartesian coordinates) and $\delta$ the wave length. The wave solutions are represented by $\dot{U}(t), \dot{\Gamma}(t), \dot{L}(t)$ and $\dot{T}(t)$.
Considering Eq. (25) together with equilibrium condition on the discontinuity surface and plastic consistency condition for gradient plasticity, the solution for drained conditions results

\[
\left( \frac{2\pi}{\delta} \right)^2 \left[ E_{ijkl} - \frac{E_{ijmn} f^*_{mn} f_{pq} E_{pqkl}}{h + h^{nl}} \right] n_i n_k \dot{U} = 0 \tag{26}
\]

being \( h^{nl} \) the generalized gradient modulus. Equation (26) leads to the expression of the acoustic tensor for non-local gradient plasticity under drained conditions

\[
Q^{nl} = E_{ijkl} - \frac{E_{ijmn} f^*_{mn} f_{pq} E_{pqkl}}{h + h^{nl}} \tag{27}
\]

The localized failure condition; \( \det(Q^{nl}) = 0 \), leads to the analysis of the spectral properties of the localization tensor. Moreover, analytical solutions of the acoustic tensor’s eigenvalues and eigenvector problem leads to explicit solutions of critical hardening/softening modulus

\[
H_{\text{crit}} = n_i E_{ijkl} f^*_{kl} P_{jm} f_{no} E_{nomq} n_q - f_{ij} E_{ijkl} f^*_{kl} \quad \text{being} \quad P = Q^{-1}
\]

as well as critical localization angles \( \theta_i \), that define the localization directions \( n \), normal to the failure surface \( S \).

4 GRADIENT- AND FRACTURE ENERGY-BASED TEMPERATURE DEPENDENT LEON-DRUCKER PRAGER CONSTITUTIVE MODEL

The fracture energy-based temperature dependent Leon-Drucker Prager (TD-LDP) constitutive model for concrete has been proposed by [14] in the framework of the thermodynamically consistent gradient-based theory for poroplastic materials.

The maximum strength criterion

\[
\Phi^*(\sigma', \tau, T) = \alpha(T) \frac{3}{2} * \tau^2 + \beta(T) m_0 \left( \frac{\sqrt{6} \star \sigma'}{\sigma'} \right) - c_0 = 0
\]

is defined in the space of effective stresses to account for the poromechanical material description

\[
\sigma' = \frac{I_1}{3} - p \quad ; \quad \tau = \sqrt{2J_2} \quad \rightarrow \quad * \sigma' = \frac{\sigma'}{f_c'} \quad ; \quad * \tau = \frac{\tau}{f_c'}
\]

being \( \sigma' \) and \( \tau \) the Haigh Westergaard effective volumetric and deviatoric stress coordinates, respectively, \( I_1 \) the first invariant of total stress tensor, \( J_2 \) the second invariant of deviatoric stress tensor, and \( f_c' \) and \( f_t' \) the uniaxial compressive and tensile strengths, respectively. The friction and cohesion of the virgin material, \( m_0 \) and \( c_0 \), are calibrated at room temperature. The temperature-dependent functions \( \alpha(T) \) and \( \beta(T) \) vary according to

\[
\alpha(T) = (1 - \gamma_1 T)^{(-1)} \quad ; \quad \beta(T) = (1 - \gamma_2 T) (1 - \gamma_1 T)^{(-1)}
\]
being $\gamma_1$ and $\gamma_2$ coefficients to be calibrated depending on the concrete quality.

To capture the diverse inelastic behaviors, one single equation-based yield surface is proposed

$$\Phi(\sigma^*, \tau^*, T, h^0, s^Q) = \alpha(T)^{3/2} \tau^* \sqrt{\frac{\sigma^*}{6}} + \eta^* \sigma^* - h^0 s^Q = 0 \quad (32)$$

Whereas the adopted plastic potential surface yields

$$\Phi^*(\sigma^*, \tau^*, T, h^0, s^Q) = \alpha(T)^{3/2} \tau^* \sqrt{\frac{\sigma^*}{6}} + \eta^* \sigma^* - h^0 s^Q = 0 \quad (33)$$

being $\eta$ the volumetric non-associativity degree which varies between $0 \leq \eta \leq 1$.

The evolution of the yield and plastic potential surfaces in pre-peak regime is controlled by the hardening dissipative stress $h^0 \leq h^0 \leq 1$, while the softening dissipative stress remains constant $s^Q = 1$. When $h^0 = 1$ the TD-LDP criterion is reached.

Softening behavior of concretes under high temperatures is related to deformation processes under increasing inhomogeneities. The strength degradation during softening regime, when $h^0 = 1$, is defined by the evolution of $s^Q$ which continuously reduces as the local and non-local decohesion process develops from its maximum value down to $s^Q = 0$.

Two parallel mechanisms contribute to the instantaneous concrete strength in softening regime: the remaining strength for further fracture development in the active cracks, which is defined by a temperature dependent fracture energy-based mechanism, and the remaining strength for further mechanical/thermal degradations in the material in between cracks. The last one is defined by a gradient-based mechanism as $s^Q_{nl} = -l_c^2 H_{nl} \nabla^2 \lambda$. The gradient thermo-plastic characteristic length $l_c$, which defines the width of the degraded material between active cracks, depends on the temperature level and acting confining pressure (see [14] for detailed explanation).

5 DISCONTINUOUS BIFURCATION SOLUTIONS FOR THE TD-LDP CONSTITUTIVE MODEL

The localization analysis of the thermodynamically consistent TD-LDP constitutive model considers the particular plane strain conditions when $\sigma_z = \nu(\sigma_x + \sigma_y)$. The following material properties are considered

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elastic modulus - E</td>
<td>19305.3 MPa</td>
</tr>
<tr>
<td>Poisson’s ratio - $\nu$</td>
<td>0.2</td>
</tr>
<tr>
<td>Uniaxial compressive strength - $f'_c$</td>
<td>22.0 MPa</td>
</tr>
<tr>
<td>Uniaxial tensile strength - $f'_t$</td>
<td>2.7 MPa</td>
</tr>
<tr>
<td>Initial internal length - $l_{c,i}$</td>
<td>25.0 mm</td>
</tr>
<tr>
<td>Maximal internal length - $l_{c,m}$</td>
<td>110.0 mm</td>
</tr>
<tr>
<td>Gradient modulus - $H^{nl}$</td>
<td>470.70 MPa</td>
</tr>
</tbody>
</table>
Figure 2: (a) Critical hardening parameter along TD-LDP maximal strength criterion and (b) Numerical localization analysis with the TD-LDP criterion at peak of the uniaxial tensile and simple shear tests.

In first place, the variation of the normalized localization indicator \( \frac{\det(Q_{nl})}{\det(Q)} \) along the maximal strength criterion defined in terms of the normalized first and second Haigh-Westergaard stress coordinates and regarding variable temperature levels from 20°C to 700°C are shown in Fig. 1-a. Transition points (TP) from ductile or diffuse failure modes to brittle or localized ones with variable temperature levels are highlighted in Fig. 1-b. It can be noted that TP move towards the low confinement and tensile regimes with increasing temperature levels.

This transitions are signalized by the appearance of null values of the normalized localization indicator corresponding to discontinuous bifurcation as can be seen in Fig. 1-a. With increasing temperature levels, model predictions realistic reproduces the displacement of the transition point towards lower confinement zones.

The evaluation of the critical hardening parameter \( H_{\text{crit}} \) has been performed along the entire failure surface in the principal stress space \( \sigma_1^*/f_c' \) and \( \sigma_2^*/f_c' \), regarding variable temperature levels from 20°C to 700°C. Figure 2-a highlights the regions where localized failure modes may occur. The increase of the temperature level influences in the yield condition and destabilizes the failure modes. The results in Fig. 2-a demonstrate that the non-local TD-LDP constitutive model signalizes discontinuous bifurcation in the pre-peak regimes of stress paths in the low confinement region. High temperatures make the situation more critical.

In first place, normalized localization indicators \( \frac{\det(Q_{nl})}{\det(Q')} = 0 \) have been numerically evaluated at the peak stresses corresponding to the simple shear, uniaxial tensile, uniaxial compression and triaxial compression tests under plane strain conditions and variable temperature levels.

As can be observed in Fig. 2-b, in case of the simple shear and the uniaxial tensile
tests, failure conditions is fulfilled in the critical directions $\theta_t = 13^\circ - 167^\circ$ and $\theta_s = 27^\circ - 153^\circ$, respectively, agreeing at 20 and 700°C.

By contrast, in the uniaxial compression test the failure mode turns ductile already, i.e. the localized failure condition is not fulfilled. Regarding Fig. 3-a, for increasing temperature levels from 20°C to 700°C, critical localization angles increase from $\theta_c = 33^\circ - 147^\circ$ to $\theta_c = 40^\circ - 140^\circ$, whereas the discontinuous bifurcation conditions are suppressed.

Finally, the influence of the increasing confining pressure in the evolution of failure modes of the TD-LDP material under plane strain conditions and variable temperature levels is numerically analyzed in Fig. 3-b. Localized failure conditions $[\det(Q^{nl})/\det(Q^e) = 0]$ have been evaluated at peak of the uniaxial and triaxial compression tests satisfying the condition $I_1/f'_c = 3$, subjected to increasing temperature levels from 20°C to 500°C. The results in Fig. 3-b demonstrate the absence of discontinuous bifurcation in the high confinement regime and, moreover, the increment of the positive definition of the localization tensor as well as of the related critical localization angle with the acting confinement and the increment of the temperature level.

6 CONCLUSIONS

The results of the localization analysis in this work demonstrate the capability of the thermodynamically consistent gradient- and fracture energy-based temperature dependent Leon-Drucker Prager (TD-LDP) constitutive model for poroplastic materials to realistically predict both brittle and ductile failure modes of concrete subject to high temperature levels when the governing stress state varies from the tensile and low con-
finement regime to the high confinement one. Analytical procedures for localized failure evaluations in thermodynamically consistent gradient and fracture energy-based materials have been applied. Numerical analysis of the condition for discontinuous bifurcation are based on the evaluation of the spectral properties of the acoustic or localization tensor through the calculation of its determinant.

The obtained results for the localization analysis under plane strain conditions demonstrate the capabilities of the constitutive model for concrete to reproduce diffuse failure modes in the medium and high confinement regimes as well as localized failure in the tensile regime and its corresponding variations with increasing temperature levels.

In conclusion, the TD-LDP material is able to reproduce the transition from brittle to ductile failure modes of concrete under plane strain conditions when the stress state varies from the tensile to the compressive regime with increasing confinement and temperature levels. The results also demonstrate the capability of the combined gradient and fracture energy-based material theory to realistically reproduce the variation of the critical localization directions with increasing temperature levels, in agreement with the concept that concretes turn more brittle when they are subjected to high temperatures.

REFERENCES


STRUCTURAL EFFICIENCY OF MULTISCALE TOPOLOGY OPTIMIZATION OF CELLULAR MATERIALS USING BI-DIRECTIONAL EVOLUTIONARY STRUCTURAL OPTIMIZATION

TAINAN K. L. CALIXTO¹, RENATO PAVANELLO²

¹Department of Computational Mechanics, Faculty of Mechanical Engineering, University of Campinas  
Rua Mendelejev 200, 13083-860, Campinas, Brazil  
tainankhalil@gmail.com

²Department of Computational Mechanics, Faculty of Mechanical Engineering, University of Campinas  
Rua Mendelejev 200, 13083-860, Campinas, Brazil  
pava@fem.unicamp.com

Key words: Topology Optimization, Multiscale design, BESO, Homogenization Method

Abstract. Topology Optimization is an active field of research with a large number of academic and industrial applications. One important application is the development of new materials considering the design in the microstructure level. In this context, it is possible to consider the design of the microscopic domain to improve the performance of the structures. In this paper, it is proposed to apply an evolutionary technique to design new cellular materials using Topology Optimization in the periodic micro scale domain. In cellular materials, the microstructure has only one material and voids. Using the homogenization theory, the material design assumes that the macrostructure material is made of periodic base cells. The macroscopic properties of the heterogeneous material are homogenized, according to the micro scale topology. This work presents an approach of the Bi-directional Evolutionary Structural Optimization (BESO) method for designing microstructures of cellular materials. The objective function is to minimize the mean compliance of structures or to maximize the natural frequency of the dynamic response for predefined macro scale problems. Numerical examples are presented to validate the effectiveness of the algorithm. The influence of the mesh refinement and the initial design in the final topology of the microstructure are discussed. Finally, a structural efficiency index is proposed for multiscale problems and some results for cellular materials are presented. The structural efficiency indexes of optimized microstructures for different macroscopic boundary conditions are analysed.
1 INTRODUCTION

The development of high performance material is a reality nowadays. This became true due to the high facility to produce new materials with 3D printers, for example. Therefore engineers are challenged to use light materials with good properties, instead of use common solid metal.

Multiscale topology optimization methods have been an important tool in the recent years to the development. In order to apply topology optimization to the microscopic model, the homogenization theory, as presented in Hassani and Hinton [2, 3], is used. The first relevant work about optimal shape design using homogenization method was developed by Bendsøe and Kikuchi [1]. More recently, Huang et al. [5] and Zuo et al. [6] has studied the multiscale topology optimization using Bi-directional Evolutionary Structural Optimization (BESO) for static and dynamic problems, respectively. In a recent work, the authors has studied the minimization of the frequency responses functions of multiscale systems composed of macro and micro phases [8].

Although the topology optimization results in more efficient structures, this is not always true in multiscale problems with periodic micro scale, due to the presence of material in inefficient location of the final structure. A comparative of the objective function ratios between the optimized micro cell and the initial filled base cell, for the same material volume, can be used to evaluate the structural efficiency. An index is proposed to evaluate the performance of the local optimal solutions.

In this work, a multiscale model and a topology optimization method will be briefly presented. In the numerical results, the optimization algorithm is applied to several examples, including static and dynamic problems. Lastly, the effectiveness of the optimized microstructures are analysed using a performance index for some different macroscopic boundary conditions.

2 MULTISCALE MODEL - HOMOGENIZATION METHOD

For a multiscale modelling, the concepts of the homogenization theory was used considering a two-scales problem. According to this theory, if a material presents low periodic inhomogeneities, the homogenized properties of the base cell can be obtained [2, 3]. Using asymptotic analysis, as a result, we can express the mechanical macroscopic properties as a function of the microstructural configuration. The Figure 1 shows the micro and macro scales considered in the homogenization theory for a structure with periodic base cell.

In order to use the homogenization theory, the microscopic scale should be much smaller than the macroscopic scale.

Using the finite element method, the multiscale model should have two different meshes, one for each scale or model. In the microscopic scale, applying the periodic boundary condition and the homogenization conditions, the equilibrium equation can be written as [2, 3]:
\begin{equation}
\int_{Y} b^T D b dY u = \int_{Y} b^T D dY
\end{equation}

where \(b\) is the strain/displacement matrix, \(D\) is the elasticity matrix of the material, \(u\) is the displacement fields in the micro scale model and \(Y\) is the volume of the base cell. At the left side of equation 1, \(k = \int_{Y} b^T D b dY\) is the stiffness matrix of the microstructure level. At the equation 1, the right side represents the external forces caused by three uniform strain fields \(\{1, 0, 0\}^T\), \(\{0, 1, 0\}^T\) and \(\{0, 0, 1\}^T\) for 2D problems [2, 3].

Finally, the homogenized elastic matrix \(D^H\) for the microstructure is calculated using the following expression:

\begin{equation}
D^H = \frac{1}{|Y|} \int_{Y} D (I - bu) dY
\end{equation}

where \(I\) is an identity matrix and \(u\) denotes the displacement fields of the unit cell caused by the uniform strain fields.

3 OPTIMIZATION PROBLEM

In this work, we realize two different types of optimization: for static and dynamic problems. The static problem aims to search the stiffest structure with a given volume of material. The mean compliance \(C\) is defined as the total strain energy of the structure. The static topology optimization problem can be written as:

Find: \(x\)

that minimize:

\(C = \frac{1}{2} F^T U\) \hspace{1cm} (3)

subject to:

\(K U = F\) \hspace{1cm} (4)

\(V^* - \sum_{j=1}^{M} V_j x_j = 0\)

\(x_j = 0\) or \(1\)
where $x_j$ is the $j$th design variable of the micro model and $V^*$ is the target volume. $U$ is the nodal displacement field vector of the macrostructure. $K$ and $F$ are respectively the global stiffness matrix and the global force vector of the macro model.

For the dynamic problem, we seek to maximize the natural frequency of the structure. In order to obtain a local optimal solution considering dynamic criteria, the topology optimization problem is formulated as below:

Find: $x$

that maximize: $\omega_k$

subject to:

$$\begin{align*}
(K - \omega_k^2 M) U_k &= 0 \\
V^* - \sum_{j=1}^{M} V_j x_j &= 0 \\
x_j &= 0 \text{ or } 1
\end{align*}$$

(5)

where $\omega_k$ is the natural frequency $k$ of the macrostructure, $U_k$ is the correspondent eigenvector $k$, $M$ and $K$ are the global mass and stiffness matrices of the macro model, respectively. Finally, $x$ describe the layout of the micro model and $V^*$ is the target volume.

The material of an element in the base cell is modelled considering an isotropic material behaviour. Its physical properties are assumed to be a function of the elemental density $x_j$. The material interpolation scheme named SIMP (Solid Isotropic Material with Penalization) [4] is adopted, and the following approximation is used:

$$D(x_j) = x_j^p D$$

where $x_j = x_{\text{min}}$ or 1

(6)

where $p$ is the exponent of penalization. In this paper, $p = 3$ is used. The soft-kill approach indicates that no element is allowed to be completely removed from the design domain ($x_{\text{min}} = 0.001$)[4].

4 SENSITIVITY ANALYSIS

The sensitivity analysis of the multiscale problem involves the derivative of the objective function in macro scale with respect to the variable of the micro scale. In this context, both models, micro and macro, must be considered in the calculation of sensitivities.

4.1 Static Problem

For this problem, the sensitivity analysis of the macrostructural mean compliance $C$ is derived from the equation 3. Using the material model done in equation 6, it can be defined as a function of the microstructural variable $x_j$, as following [5]:

$$\alpha_j^s = -\frac{1}{p} \frac{dC}{dx_j}$$

$$\alpha_j^s = \frac{1}{2p|Y|} \sum_{i=1}^{M} U_{k,i}^T \left\{ \int_{V_i} B^T \left[ \int_Y (I - bu)^T \left( \frac{\partial D}{\partial x_j} \right) (I - bu) dY \right] B dV_i \right\} U_{k,i}$$

(7)
where $M$ is the number of elements of the macroscopic mesh and the term between braces represents the derivative of the elemental stiffness matrix ($\frac{\partial K_e}{\partial x_j}$).

### 4.2 Dynamic Problem

The sensitivity analysis for the dynamic problem starts with the Rayleigh quotient, that is derived from the equilibrium equation of the macro model (5). The eigenvalue $\omega_k$, defined as a function of the macroscopic parameters only, can be written as:

$$\omega_k^2 = \frac{U_k^T K U_k}{U_k^T M U_k}$$

(8)

The derivative of the macrostructural natural frequency from the Rayleigh quotient (8) with respect to the design variable $x_j$ of the micro model is calculated by:

$$\frac{\partial \omega_k}{\partial x_j} = \frac{1}{2\omega_k} \left[ U_k^T \left( \frac{\partial K}{\partial x_j} - \omega^2 \frac{\partial M}{\partial x_j} \right) U_k \right]$$

$$\frac{\partial \omega_k}{\partial x_j} = \frac{1}{2\omega_k} \sum_{i=1}^{M} U_{k,i}^T \frac{\partial K_e}{\partial x_j} U_{k,i} - \frac{\omega_k}{2} \sum_{i=1}^{M} U_{k,i}^T \frac{\partial M_e}{\partial x_j} U_{k,i}$$

(9)

Considering that, in equation 9, the change of the elemental mass matrix in the macro model due to each element in the micro model is constant [6], the sensitivity of the elements in the microstructure can be obtained as the following equation:

$$\alpha^d_j = \frac{\partial \omega_k}{\partial x_j} + \frac{\omega_k}{2} \sum_{i=1}^{M} U_{k,i}^T \frac{\partial M_e}{\partial x_j} U_{k,i}$$

$$\alpha^d_j = \frac{1}{2|Y|\omega_k} \sum_{i=1}^{M} U_{k,i}^T \left\{ \int_{V_i} \mathbf{B}^T \left[ \int_{Y} (\mathbf{I} - \mathbf{b}u)^T \left( \frac{\partial \mathbf{D}}{\partial x_j} \right) (\mathbf{I} - \mathbf{b}u) dY \right] \mathbf{B} dV_i \right\} U_{k,i}$$

(10)

It can be noticed that the static and dynamic sensitivity equations (7 and 10, respectively) are similar, due to fact that, for both models, only stiffness has been considered.

### 5 BI-DIRECTIONAL EVOLUTIONARY STRUCTURAL OPTIMIZATION PROCEDURE

The optimization procedure applied in this work is the BESO method. It consists in gradually removes and adds material on the finite element mesh. The evolutionary optimization uses a discrete design variable, which means that the material can only be void or solid. A mesh-independent filter was used to avoid numerical instabilities. A history averaging scheme is used to improve the convergence of the solutions. The method details are presented by Huang and Xie [4].

The summary of the evolutionary optimization procedure is presented in the flowchart of Figure 2. The algorithm was implemented using MATLAB and the numerical results shall be presented in the next section.
Define macrostructure domain, target volume $V^*$ and BESO parameters

Micro model: Carry out FE analysis and calculate the homogenized material properties matrix for the base cell

Substitute $D^h$ to elements in the macro scale and solve the problem

Calculate the element sensitivities in the micro model

Process element sensitivities with filtering and history-averaging

Update the elements of the design according to the element sensitivities and the current target volume in the micro model

Volume constraint satisfied?

Yes

Convergence criterion satisfied?

Yes

Local optimal solution

No

No

Figure 2: BESO procedure for multiscale optimization problem

The nomenclature of the BESO parameters used in this work is equivalent to the presented in [4], [7] and [9].

6 NUMERICAL RESULTS AND DISCUSSIONS

In this section, we will show some results of the multiscale topology optimization for the static and dynamic cases. In these examples, only the microscopic scale will be designed. The efficiency of the base cell optimization shall be analysed, for different macroscopic boundary conditions. In all analyses, the properties of the isotropic solid material used are Young’s modulus $E = 1.0$, Poisson’s ratio $\nu = 0.3$ and density $\rho = 1.0$. For the finite element mesh, quadrilateral elements were used. The base cell is discretized into 100 x 100 four-node quadrilateral finite elements. The same element type with size 1 x 1 is assigned for all macrostructures. The BESO parameters used in all simulations are: $ER = 2\%$, $r_{\text{min}} = 4$, $AR_{\text{max}} = 2\%$ and $\tau = 0.1\%$. To start the optimization procedure, the initial design of the microscopic mesh is full with solid material except for four elements at the center of the base cell with no material (voids).

6.1 Multiscale Topology Optimization

Initially, the algorithm has been validated for several numerical examples. For static analysis, the applied force is unitary ($F = 1$ N) and its position is shown at the following figures. The objective volume $V^*$ applied for this section is 40% of material.
The first example is the cantilever beam with dimensions 40 x 20. The numerical results is shown at Figure 3, where 3(a) is the macrostructure model with boundary conditions; 3(b) and 3(c) is the optimized topology of cellular material and the 3x3 unit cells for the static problem respectively; 3(d) and 3(e) is the optimized topology of cellular material and the 3x3 unit cells for the fundamental frequency maximization of the system. The next two figures will follow this same pattern of organization.

![Figure 3](image)

**Figure 3**: Optimized topologies of cellular materials for Cantilever beam with $L = 40$ and $H = 20$ (a). Static case: base cell (b), 3x3 unit cells (c); Dynamic case: base cell (d), 3x3 unit cells (e).

The results indicates that the structural stiffness of x-direction for both unit cells are higher than that of y-direction. This was expected since the normal stress caused by bending is much higher than the shear stress for a cantilever beam with these dimensions.

For a MBB beam with dimensions 80 x 20, the results are shown at Figure 4. To solve this problem, a symmetry condition can be imposed and only half of the macroscopic domain is used. In this paper, the left side of the structure was optimized, so the microstructure of the right side should be symmetric.

For static problem, the results presents in 3(b)/3(c) and 4(b)/4(c) are compared with the results presented in [5], showing a good agreement.

Finally, for a double-clamped beam 5(a) with dimensions 40 x 20, Figure 5 presents the results for static (b,c) and dynamic (d,e) cases.

It can been noted that, for all different types of macroscopic boundary conditions, the topologies obtained for static and dynamic problems are similar. This is explained by the fact that, for the dynamic analysis, the mode shapes of the first natural frequency are similar to the displacement field in the static solution.
6.2 Structural Efficiency Analysis

In order to verify the effectiveness of this kind of optimization procedure, a preliminary structural efficiency index is proposed. For a given target volume $V^*$, the objective func-
The material reduction will vary between 0% and 70%.

For the static multiscale analysis, the objective is to evaluate the mean compliance ratio between the reduced structure and the one obtained from the optimization problem, as presented in Figure 6. For the dynamic problems, the objective is the first fundamental frequency ratio obtained from the multiscale optimization result and from the reduced structure. Both analysis parameters are presented in equation 11.

\[
\gamma_s = \frac{C_{\text{red}}(V^*)}{C_{\text{opt}}(V^*)} > 1 \quad \gamma_d = \frac{\omega_{\text{opt}}(V^*)}{\omega_{\text{red}}(V^*)} > 1
\]  

where the structural efficiency for the static problem represented by \( \gamma_s \) should be greater than 1, since the problem is to minimize the mean compliance of the structure. As the dynamic problem consists in the first fundamental frequency maximization, the ratio \( \gamma_d \) is the inverse of the static one and it should also be greater than unity.

First, the Figure 7 shows the results for a cantilever beam. It can be seen that, for all target volumes calculated, the ratio \( \gamma_s \) and \( \gamma_d \) are both greater than 1, in other words, the multiscale optimization always improves the structure for static and dynamic analyses.

The Figure 8 shows the results for a MBB beam. Until the material reduction of 25%, the efficiency \( \gamma_s \) and \( \gamma_d \) are less than 1, so the optimization results is not better than the reduction of the height of the structure (for static and dynamic).

Finally, the results of the efficiency index for a double-clamped constraint is shown in Figure 9. For static analysis, the mean compliance ratio is less than 1 until almost 40% of the material reduction. It means that the multiscale optimization should be used for big values of material reduction (50% or plus). For dynamic analysis, the double-clamped structure is similar to the MBB beam, and the optimal topology should be used above 30% of volume reduction.
It is important to notice that, for this efficiency analysis, the results presented in this paper evaluate only material cost. It means that no fabrication cost was considered for all comparison.

7 CONCLUSIONS

This paper has analysed the multiscale topology optimization which designs a microstructure of cellular materials for a given volume fraction so that the macrostructure has the maximum stiffness and maximum first natural frequency. The algorithm was
Numerical examples were presented for different material reduction for cantilever, MBB and double-clamped beam. The structural efficiency was analysed for these three type of boundary conditions. It could be seen that, for low material reduction at the base cell, the respective topologies obtained are similar for static and dynamic problems (as expected) and the structural efficiency is very low or even less than 1. For high material reduction, the structural efficiency is high and the multiscale topology optimization of cellular materials proved to be a good alternative for the development of new materials.

ACKNOWLEDGEMENTS

This work was supported by the National Council for Scientific and Technological Development (CNPq) [grant 155808/2013-3] and the Sao Paulo Research Foundation (FAPESP) [grant 2013/08293-7].

REFERENCES


CONTINUOUS MICROPLANE THEORY AND INTERFACE APPROACH FOR FAILURE FORM ANALYSIS OF STEEL FIBER REINFORCED CONCRETE

SONIA M. VRECH\textsuperscript{1}, ANTONIO CAGGIANO\textsuperscript{2} AND GUILLERMO ETSE\textsuperscript{1,2}

\textsuperscript{1}CONICET - University of Tucumán
Av. Independencia 1800. 4000 - San Miguel de Tucumán. Argentina
svrech@herrera.unt.edu.ar

\textsuperscript{2}CONICET - University of Buenos Aires
Av. Las Heras 2214 - C1127AAR - Buenos Aires. Argentina
acaggiano@fi.uba.ar
getse@herrera.unt.edu.ar

Key words: Discontinuous bifurcation, microplanes, SFRC, interfaces, fracture

Abstract. A fracture energy-based microplane constitutive theory for steel fiber reinforced concrete (SFRC) is presented to evaluate the properties of the discontinuous bifurcation condition under different scenarios of stress states, fiber contents and directions. The constitutive model considers a CAP-cone yield surface of $C^1$ continuity at the microplane level. Its evolution in the post-peak regime is described by means of a fracture energy-based work softening which is defined differently for mode I and II type of failure. The effect on the post-peak ductility introduced by the fiber content is also taken into account, while the directional properties of the steel fibers are considered through the relative directions between microplanes and fibers.

The main objective of the discontinuous bifurcation analysis proposed in this work, is to evaluate the capabilities of the microplane theory to capture the directional enrichment provided by steel fibers to the ductility and also, to reproduce the particular microcrack directions which (in the framework of the smeared crack approach) are mathematically represented by the spectral properties of the critical localization tensor. Firstly, the localized failure features in the form of discontinuous bifurcation of SFRC are identified by means of numerical analysis. Mono- and multidirectional fiber distributions and different steel fiber contents are considered in the localized failure analysis to be performed on stress states corresponding to critical strengths of SFRC under both uniaxial and biaxial tension and compression. Then, microplane model are compared with the FE predictions obtained with the interface model for FRC previously proposed by the authors. In the last case, the crack evolutions and their directions are explicitly described throughout the so-called discrete crack approach.
1 INTRODUCTION

Localized failure mechanisms on quasi-brittle materials, understood as the formation of restricted failure zones with high concentration of deformations while the rest of the structure might even exhibit unloading, depend on their acting stresses as well as the mechanical and chemical features which they are submitted and have experimentally been observed by a. o. [1, 2].

While the localization of deformations in terms of discontinuous bifurcation for plain concrete has extensively been studied, see a.o. [3, 4], there are not yet similar analysis related to fiber reinforced concrete (FRC). In this case it becomes necessary to distinguish failure mechanisms that characterize tension, compression and shear regimes regarding different fiber contents as well as fiber directions.

In this work, a novel constitutive formulation for SFRC is formulated in the framework of the Mixture Theory by [5] and based on the microplane model. At microplanes level, it is described in terms of normal and shear stresses and their related strains. An hyperbolic yield surface has been adopted while a CAP-cone one of $C^1$ continuity describes the constitutive behavior in the high confinement regime. Their evolution in the post-peak regime is described by means of a fracture energy-based work softening which is defined differently for mode I and II type of failure. Fiber effect on the composite failure behavior is taken into account through both a bond-slip formulation and a dowel model depending on the relative orientations between fibers and microplanes.

Then, the capabilities of the microplane theory to capture the directional enrichment provided by steel fibers to the ductility and to reproduce the particular microcrack directions are evaluate and compared against FE predictions performed by the authors in the framework of the discrete approach [6].

The thermodynamically consistent microplane-based elasto-plasticity theory is summarized in Section 2 while the analytical solution for localized failure in terms of discontinuous bifurcation is developed in Section 3. Then, in Section 4 the adopted composite constitutive formulation for SFRC is described. Finally, Section 5 shows the localized failure analysis for SFRC and later, the concluding remarks are highlighted in Section 6.

2 THERMODYNAMICALLY CONSISTENT MICROPLANE-BASED ELASTOPLASTICITY

The thermodynamically consistent microplane-based elasto-plasticity theory for the derivation of macroscopic stresses and equilibrium equations in the case of isotropic plasticity have been developed by [7] and [8].

Assuming kinematic constraints, scalar volumetric strain and tangential strain vector at microplane level ($\varepsilon_V$ and $\varepsilon_T$, respectively) are computed by means of the following relationships

$$\varepsilon_V = V : \varepsilon \quad , \quad \varepsilon_T = T : \varepsilon$$

(1)
being $\varepsilon$ the macroscopic strain tensor projected on a microplane of normal direction $n$. The projection tensors are defined as

$$V = \frac{1}{3}I, \quad T = n \cdot I^{ym} - n \otimes n$$

being $I^{ym}$ the symmetric part of the fourth-order identity tensor.

Thus, the strain vector at microplane level results

$$t_\varepsilon = \varepsilon_V n + \varepsilon_T.$$  (3)

Assuming the macro free-energy potential as the integral of the micro free-energy on a spherical region of unit volume $\Omega$, the following micro-macro free-energy relationship is proposed

$$\psi_{mac} = \frac{3}{4\pi} \int_\Omega \psi_{mic} d\Omega$$  (4)

being $\psi_{mic}(\varepsilon_V, \varepsilon_T, \kappa)$ the free-energy potential at microplane level, expressed in terms of the strain components and the scalar internal variable $\kappa$. Assuming small strains and considering the additive decomposition into elastic and plastic parts of the macroscopic strain tensor, the microscopic strain components are expressed as

$$\varepsilon_V = \varepsilon^e_V + \varepsilon^p_V, \quad \varepsilon_T = \varepsilon^e_T + \varepsilon^p_T.$$  (5)

Then, the constitutive micro-stresses and their rates are computed as

$$\sigma_V = \frac{\partial \psi_{mic}}{\partial \varepsilon_V} \rightarrow \dot{\sigma}_V = E_V [\dot{\varepsilon}_V - \dot{\varepsilon}_V^p]$$

$$\sigma_T = \frac{\partial \psi_{mic}}{\partial \varepsilon_T} \rightarrow \dot{\sigma}_T = E_T [\dot{\varepsilon}_T - \dot{\varepsilon}_T^p]$$  (6)

while the dissipative stresses and their rates can be computed at microplane level as

$$\phi_{mic} = \frac{\partial \psi_{mic}}{\partial \kappa} \rightarrow \dot{\phi}_{mic} = \dot{\kappa}$$  (7)

being $\dot{\kappa}$ the hardening/softening modulus.

As in case of macroscopic plasticity, both yield and plastic potential surfaces are set as

$$\Phi^{mic}(\sigma_V, \sigma_T, \phi_{mic}) \leq 0 \quad \text{with} \quad \nu_V = \frac{\partial \Phi^{mic}}{\partial \sigma_V} \quad \text{and} \quad \nu_T = \frac{\partial \Phi^{mic}}{\partial \sigma_T}$$

$$\Phi^{*mic}(\sigma_V, \sigma_T, \phi_{mic}) \leq 0 \quad \text{with} \quad \mu_V = \frac{\partial \Phi^{*mic}}{\partial \sigma_V} \quad \text{and} \quad \mu_T = \frac{\partial \Phi^{*mic}}{\partial \sigma_T}$$  (8)

and the evolution of the plastic strain components yields
\[ \dot{\varepsilon}_V^p = \dot{\lambda} \mu_V, \quad \dot{\varepsilon}_T^p = \dot{\lambda} \mu_T, \quad \kappa = \dot{\lambda}. \quad (9) \]

Moreover, the Kuhn-Tucker and consistency conditions must be satisfied
\[ \Phi^{mic} \leq 0, \quad \dot{\lambda} \geq 0, \quad \Phi^{mic} \dot{\lambda} = 0 \quad \text{and} \quad \dot{\Phi}^{mic} \dot{\lambda} = 0 \quad (10) \]

The homogenization of the microplanes energy of Eq. (4) leads to the definition of the macroscopic stress tensor as
\[ \sigma = \frac{\partial \psi^{mac}}{\partial \varepsilon} = \frac{3}{4\pi} \int_{\Omega} \left[ V \sigma_V + T^T \cdot \sigma_T \right] d\Omega. \quad (11) \]

The analytical evaluation of this integral can be solved by numerical integration techniques proposed by [9]. Thus, Eq. (11) can be rewritten as
\[ \sigma \approx \sum_{I=1}^{n_{mic}} \left[ V^I \sigma_V^I + T^{TI} \cdot \sigma_T^I \right] w^I \quad (12) \]

where the superscript \( I \) denotes the \( I \)th material direction and \( w^I \) the corresponding weight coefficients. The number of microplanes \( n_{mic} \) that ensures accurate approximations is 42.

The macroscopic tangent operator can be analogously obtained as
\[ E^{ep} = \frac{d\sigma}{d\varepsilon} = \frac{3}{4\pi} \int_{\Omega} \left[ V \otimes \frac{d\sigma}{d\varepsilon_V} + T^T \cdot \frac{d\sigma}{d\varepsilon_T} \right] d\Omega. \quad (13) \]

resulting
\[ E^{ep} = E^e - \frac{3}{4\pi} \int_{\Omega} \frac{1}{h} \left[ E^e_V V \mu_V + E^e_T T^T \cdot \mu_T \right] \otimes \left[ \nu_V E^e_V + \nu_T \cdot T^T E^e_T \right] d\Omega \quad (14) \]

with the elastic macroscopic tangent operator computed as
\[ E^e = \frac{3}{4\pi} \int_{\Omega} E^e_V V \otimes V + E^e_T T^T \cdot T d\Omega. \quad (15) \]

3 ANALYTICAL SOLUTION FOR LOCALIZED FAILURE IN MICROPLANE-BASED ELASTO-PLASTICITY

In the framework of the smeared crack approach, localized failure modes are related to discontinuous bifurcations of the equilibrium path, and lead to lost of ellipticity of the equations that govern the static equilibrium problem. The inhomogeneous or localized deformation field exhibits a plane of discontinuity that can be identified by means of the eigenvalue problem of the acoustic or localization tensor, see [10]. Analytical solutions for the discontinuous bifurcation condition conduce to the macroscopic localization condition
In case of microplane-based plasticity, the acoustic tensor is expressed as

\[
Q^{ep} = Q^e - \frac{3}{4\pi} \int_{\Omega} \frac{a^* \otimes a}{h} d\Omega
\]  

(17)

with the traction vectors computed as

\[
\begin{align*}
\mathbf{a} &= [\nu_V E_V^e + \nu_T \cdot T_E_T^e] \cdot \mathbf{n}, \\
\mathbf{a}^* &= \mathbf{n} \cdot [E_V^e V \mu_V + E_T^e T \cdot \mu_T].
\end{align*}
\]  

(18)

Analytical solutions of the acoustic tensor’s eigenvalues and eigenvector problem lead to explicit solutions of critical hardening/softening modulus

\[
\bar{H}_c = \mathbf{a} \cdot [Q^{ep}]^{-1} \cdot \mathbf{a}^* - \nu : E^e : \mu = 0
\]  

(19)

as well as critical localization angles \( \theta_i \), that define the localization directions \( \mathbf{n} \), normal to the failure surface \( S \).

Due to the complex structure of the acoustic tensor for microplane-based plasticity in Eq. (17), the analytical assessment is not easy. Instead, numerical solutions must be applied and Eq. (17) can be rewritten as

\[
Q^{ep} \approx Q^e - \sum_{l=1}^{n_{\text{mic}}} \left[ \frac{a^{*I} \otimes a^{I}}{h^I} \right] w^I.
\]  

(20)
Figure 2: Numerical localization analysis with the microplane-based criterion at peak of the simple shear tests of SFRC with 8% fiber content.

4 COMPOSITE CONSTITUTIVE FORMULATION FOR SFRC

In this section, the constitutive formulation for SFRC based on microplanes and Mixture Theories [5] is described. On this basis, each infinitesimal volume of the composite is characterized by the same amount and proportion of its constituents (mortar matrix and randomly oriented fiber reinforcements for SFRC). Consequently, the composite stress vector is defined as the weighted sum (in terms of the volume fraction) of the constituent stresses

\[ t_\sigma = w_m \sigma^m + \sum_{f=1}^{n_f} w_f [\sigma (\varepsilon_N) n_f + \tau (\varepsilon_T) t_f] \]  

being \( \omega_m \) and \( \omega_f \) the weighting functions depending on the volume fraction of each constituent, with \( m \) and \( f \) indicating mortar and fiber, respectively. The mortar stress vector is computed as \( \sigma^m = [\sigma_N \ \sigma_T] \); \( \sigma \) and \( \tau \) mean the bond-slip and dowel stresses of the single fibers, related to their axial and tangential strains, \( \varepsilon_N \) and \( \varepsilon_T \); while \( n_f \) and \( t_f \) are the vectors in the fiber and its orthogonal directions, respectively.

4.1 Failure surfaces evolution for plain mortar

The maximum strength criterion of the aforementioned fracture energy-based plasticity formulation for plain mortar combines the three-parameter hyperbolic strength function by [11], in terms of the initial tensile strength \( \chi_0 \), cohesion \( c_0 \) and the internal friction angle \( \varphi_0 \); with an elliptical CAP surface to more accurately compute mortar peak capacity in the high confinement regime. There are expressed as

\[ \Phi_{hyp} = \sigma_T^2 - [c_0 - \sigma_N \tan(\varphi_0)]^2 + [c_0 - \chi_0 \tan(\varphi_0)]^2 \text{ if } \sigma_{C,0} \leq \sigma_N \leq \chi_0 \]  

254
\[ \Phi_{\text{CAP}} = \sigma_T^2 - \left[ \frac{(\sigma_{L,0} - \sigma_{A,0})^2 - (\sigma_N - \sigma_{A,0})^2}{\tau_0^2} \right] = 0 \quad \text{if} \quad \sigma_N < \sigma_{C,0} \quad (23) \]

being \( \sigma_T = \| \sigma_T \| \), \( \sigma_{A,0} \), \( \sigma_{L,0} \) and \( \tau_0 \) are material parameters of the CAP. The limit strength \( \sigma_{C,0} \) corresponds to the particular value of \( \sigma_N \) for which the equality and continuity conditions for \( \Phi_{\text{hyp}} \) and \( \Phi_{\text{CAP}} \) take place.

The plastic flow rule, computed as
\[ \dot{\varepsilon}^p = \left[ \begin{array}{c} \dot{\varepsilon}_N^p \\ \dot{\varepsilon}_T^p \end{array} \right] = \lambda A \cdot \left( \begin{array}{c} \nu_N \\ \nu_T \end{array} \right) \quad (24) \]

being
\[ \nu_N = \frac{\partial \Phi}{\partial \sigma_N} , \quad \nu_T = \frac{\partial \Phi}{\partial \sigma_T} \quad (25) \]
can be subdivided in five zones according to the confinement level, characterized by particular expressions of the matrix \( A \).

The post-cracking behavior is controlled by the rate of the plastic work spent, named \( \dot{\tilde{w}} \), during three possible post-peak failure mechanisms: mode I, shear or mode II and pure compaction.

For the sake of brevity, the complete description of the models are omitted. Further details can be found in [12].

4.2 Failure surfaces evolution for steel fibers

Steel fiber actions, developed on active opened cracks, offer bridging effects on the overall SFRC post-peak behavior. On the one hand, the bond-slip mechanism between fibers...
and concrete matrix is treated as an axial (tensile) fiber stress acting at microplane level. On the other hand, a dowel effect is considered to model the shear transfer mechanism which develops in case of fibers crossing active cracks while they are subjected to stresses in the normal direction to fibers.

The complete formulations of both, the bond-slip and dowel action, have been detailed in [12].

5 SOLUTION FOR DISCONTINUOUS BIFURCATION ANALYSIS OF SFRC

In this section, numerical localization analysis of SFRC have been performed regarding the microplane-based formulation developed in previous sections. Based on the 2D formulation of the microplane-based elasto-plasticity by [13], 48 microplanes have been considered for the integration of stresses at each material point. Plane strain state has been considered, regarding the particular condition when \( \sigma_z = \nu(\sigma_x + \sigma_y) \).

The evaluation of localization conditions signaled by the appearance of null values of the normalized determinant of the acoustic tensor corresponding to discontinuous bifurcation \( [\text{det}(Q^{ep})/\text{det}(Q^e)=0] \), has been carried out at the peak stresses corresponding to the simple shear, uniaxial tensile, uniaxial compression and biaxial tests.

The model parameters are adjusted according the experimental data given by [14], being \( E_c = 39.5\, GPa \) and \( \nu = 0.20 \), \( \tan(\phi_0) = 0.6 \), \( \chi_0 = 4.0\, MPa \), \( c_0 = 7.0\, MPa \), \( \sigma_{C,0} = -7.0\, MPa \), \( \tau_0 = 0.25 \), \( G^I_f = 0.12\, N/mm \) and \( G^{IIa}_f = 1.2\, N/mm \). While, those corresponding with the fiber-concrete interactions are: \( E_s = 200.0\, GPa \), \( \nu = 0.30 \), \( \sigma_{y,s} = 1.2\, MPa \) and \( \tau_{y,f} = 2.35\, MPa \).

As can be observed in Figs. (1) and (2), in case of the uniaxial tensile and the simple shear tests, failure conditions are fulfilled in the critical directions \( \theta = 0^\circ \) - \( 180^\circ \) and \( \theta = 45^\circ \).
- $135^\circ$, respectively, agreeing for all fiber directions. Figure (3) indicates the position of plastic microplanes for the both cases. In the first case, it corresponds to microplanes located between $0^\circ$ and $15^\circ$ on each hemisphere and in the second case, to those between $30^\circ$ and $80^\circ$.

Under biaxial and uniaxial compression tests, the localized failure condition is also fulfilled. Regarding Figs. (4) and (5), the numerical localization condition is achieved at $\theta = 48^\circ - 132^\circ$ and $\theta = 45^\circ - 135^\circ$, respectively. Plastic microplanes are located at $30^\circ$ and at $85^\circ$ in the case of biaxial test, whereas in the case of uniaxial compression, between $0^\circ$ and $15^\circ$ in each hemisphere, see Fig. (6).

The sensitivity of the failure behaviour based on the microplane theory regarding the orientation of fibers in the cementitious matrix has been evaluated. To this end, four different fiber contents, 0, 3, 8 and 20% (theoretical case) with isotropical distribution have been considered in the uniaxial tensile test. The results in terms of numerical localization analysis are shown in Fig. (7). It can be observed, critical localization directions remain unchanged with increasing fiber contents.

Finally, the capabilities of the microplane theory to reproduce the particular microcrack directions are compared against FE predictions performed by the authors [12, 6] in the frame work of the discrete approach, with the mesoscale interface model for SFRC.

The following examples serve to validate the results obtained in this section. The case of SFRC three-point bending problem with centrical notch by [15] is considered. Crack paths in tensile region reproduces a localization angle perpendicular to the load direction, see Fig. (8). In second place, the shear tests on SFRC specimens is evaluated. As can be observed in Fig. (9), a critical localization direction at $45^\circ$ is reproduced.
6 CONCLUSIONS

In this work, an elasto-plastic microplane constitutive model aimed at predicting the failure behavior of steel fiber reinforced Concrete (SFRC) has been described. The constitutive formulation considers the well-known Mixture Theory to simulate the combined bridging interactions of fibers in concrete cracks, i.e. fiber-to-concrete bond-slip as well as dowel mechanisms.

Numerical analysis of the condition for discontinuous bifurcation, based on the evaluation of the spectral properties of the acoustic or localization tensor through the calculation of its determinant, has been applied.

The obtained results for the localization analysis under plane strain conditions demonstrate the capabilities of the constitutive model for concrete to reproduce localized failure. The numerical results also demonstrate the capabilities of the proposed constitutive theory to capture the directional orientation of the steel fiber reinforcements embedded in the cementitious matrix.

REFERENCES


Figure 7: Numerical localization analysis with the microplane-based criterion at peak of the uniaxial tension tests of SFRC with variable fiber contents: 0, 3, 8 and 20%.

Figure 8: SFRC three-point bending problem with centrical notch by [15], reproduced by [6].


Figure 9: Shear tests on SFRC, reproduced by [6].


PROPOSAL OF AN INTERPOLATION FUNCTION BETWEEN THE COMPRESSIVE AND TENSILE MERIDIANS OF FAILURE AND YIELDING CONCRETE SURFACES BASED ON BEZIER CURVES

PAULA FOLINO¹ AND DAMIÁN SMIOVICH²

¹ University of Buenos Aires, Faculty of Engineering, Laboratory of Numerical Methods in Engineering (LMNI) Institute of Science Technology and Engineering "Hilario Fernández Long" (INTECIN) Av. Las Heras 2214, C1127AAV, Ciudad Autónoma de Buenos Aires, Argentina pfolino@fi.uba.ar

² University of Buenos Aires, Faculty of Engineering, Laboratory of Numerical Methods in Engineering (LMNI) Av. Las Heras 2214, C1127AAV, Ciudad Autónoma de Buenos Aires, Argentina dsmilovich@fi.uba.ar

Key words: Failure surface, Deviatoric shape, Bezier curves, Biaxial stresses state

Abstract. This paper is related with cohesive frictional materials like concrete. It is well known that failure and mechanical behaviour of such materials depend on all the three invariants, and therefore yielding and failure surfaces involving a circular deviatoric shape, neglecting the incidence of the third invariant, cannot accurately represent the main features observed in multiaxial experimental tests, particularly when low confinement levels and load scenarios leading to different Lode angles are considered. Several functions have been proposed in the literature to predict the deviatoric shape of the failure surface of cohesive frictional materials. Most of them present a lack of smoothness and therefore, are not convenient for numerical implementations. Other proposals, present a C1 continuity type leading to a deviatoric shape similar to a triangle with rounded corners. Nevertheless, the complexity involved in numerical approaches when one of these deviatoric functions is considered, usually discourages it application, particularly in the case of non local continuum formulations like gradient based plasticity models and in multiscale approaches, enough complex even without considering a non circular deviatoric shape of the failure and yielding surfaces.

In this paper the suitability of Bezier curves to represent an appropriate deviatoric shape to be considered in the failure and yielding surfaces of concrete like materials is presented. A full derivation of a numerical approach aiming to obtain an interpolation function between the compressive and tensile meridians is reported. Finally, a critical discussion about the convenience or not of using these polynomials is addressed.
1 INTRODUCTION

Concrete is a complex material characterized by a mechanical behavior governed, among other influence factors, by the stresses state. Similarly to the case of rocks, it presents a compressive strength much higher than its tensile one. It also presents a completely different behavior under these extreme uniaxial stress states. While under uniaxial compression the material behaves in a way that can be classified as quasi ductile, under uniaxial tension this behavior drastically turns to brittle. On the contrary, under triaxial compression it almost turns to ductile, while an increase in strength is observed, which value depends on the confinement level.\(^1\)

In summary, concrete failure and mechanical behavior strongly depend on the three stress invariants \(I_1, J_2\) and \(J_3\), being \(I_1 = \delta_{ij} \sigma_{ij}\) the first invariant of the stress tensor \(\sigma\); \(J_2 = s_{ij} \frac{s_{ij}}{2}\) the second invariant of the deviatoric stress tensor \(s\); and \(J_3 = s_{ij} s_{jk} s_{ki}/3\) the third invariant of the deviatoric stress tensor. Therefore, an accurate prediction of concrete failure and mechanical behavior cannot be done neglecting one or more of the three stress invariants\(^2\). Regarding this dependence, usually Haigh Westergaard stresses space is used as coordinate system in concrete constitutive modeling, with coordinates \(\xi, \rho\); and the Lode angle \(\theta\), defined as

\[
\xi = \frac{I_1}{\sqrt{3}} \quad ; \quad \rho = \sqrt{2J_2} \quad ; \quad \cos(3\theta) = \frac{3\sqrt{3}}{2} \frac{J_3}{J_2^{3/2}}
\]  

Certainly, among the three stress invariants, \(J_3\) is the most difficult to be included in constitutive models regarding the numerical complexity that it introduces in the numerical implementation. Therefore, most of the failure criteria and constitutive models for concrete, neglect it influence. Nevertheless, several research works in the literature have demonstrated the important incidence of \(J_3\) not only on determining concrete peak stress\(^2\) but also on localization properties and on structural behavior\(^{3,4}\).

In general, concrete maximum strength and yielding 3-D surfaces are based on the following process: In first place, two functions defining the compressive and tensile meridians are proposed, being \(\theta = 60^\circ\) and \(0^\circ\), respectively. Therefore, those functions only depend on coordinates \(\xi\) and \(\rho\) and can be observed on a meridian view in the Haigh Westergaard stresses space. Only once the main meridians are defined, a function defining the shape of the deviatoric view, or in other words the variation of the second Haigh Westergaard coordinate \(\rho\) with the Lode angle \(\theta\) is proposed. In the following, this function will be called interpolation function between the compressive and tensile meridians, or just interpolation function.

Several interpolation functions have been proposed in the literature. Most of them present a lack of smoothness and therefore, are not convenient from the point of view of the numerical implementation\(^{5,6}\). Among the proposals considering a \(C_1\) continuity type, two approaches must be mentioned. On the one hand, the proposal by Ottosen (1977)\(^7\) and on the other hand, the elliptical interpolation proposed by Willam and Warnke (1974)\(^8\).
both leading to a deviatoric shape similar to a triangle with rounded corners. Since these proposals were presented, no new deviatoric function has been addressed. The latter has extensively been used in different concrete failure criteria and constitutive models in the literature\textsuperscript{[4,9–12]}. Nevertheless, it has been demonstrated that the elliptical interpolation fails to accurately predict biaxial stress states\textsuperscript{[2,13]}.

The aim of this work is to propose a new function for the interpolation between the compressive and tensile meridians based on Bezier curves. The suitability of these parametric curves to represent an appropriate deviatoric shape to be considered in the failure and yielding surfaces of concrete-like materials is analyzed. Finally, a critical discussion about the convenience or not of using these polynomials is addressed.

2 CONCRETE DEVIATORIC SHAPE

In Fig. 1 typical concrete maximum strength compressive and tensile meridians are depicted, corresponding to different concrete types and based on experimental results obtained from the literature\textsuperscript{[14–16]}. It can be seen that compressive meridian depends on concrete quality while tensile meridian almost remain the same for all the concrete types that are presented in the figure. It can also be observed that both meridians are nonlinear and that the eccentricity, defined as $e = \rho_t / \rho_c$, being $\rho_t$ and $\rho_c$ the maximum shear strengths in the tensile and compressive meridians, respectively, increase with confinement, but being always $e < 1$. Then, it can be concluded that the interpolation function between the compressive and tensile meridians must be the function defining intermediate values of $\rho$ between two extreme values, $\rho_t$ and $\rho_c$, being the ratio $e < 1$.

The proposal by Willam and Warnke (1974)\textsuperscript{[8]} is based on three ellipses that can be
seen in Fig. 2 represented by the following equation

\[ r \rho = \rho_c \]  

(2)

being \( r \) the interpolation or ellipticity factor defined as

\[ r = \frac{4(1 - e^2) \cos^2 \theta + (2e - 1)^2}{2(1 - e^2) \cos \theta + (2e - 1)\sqrt{4(1 - e^2) \cos^2 \theta + 5e^2 - 4e}} \]  

(3)

3 BEZIER CURVES

Bezler curves are defined by functions relating a set of given points, denoted as control points \( P_i \), by blending functions in terms of a parameter \( 0 \leq t \leq 1 \), and are represented by the following equation

\[ B(t) = \sum_{i=0}^{n} \frac{n!}{t!(n-i)!} (1-t)^{n-i} t^i P_i \]  

(4)

being \( n \) an integer number and \( i = 0, 1, ..., n \). Therefore, quadratic Bezler curves have three control points and are defined by

\[ B(t) = (1-t)^2 P_o + 2t (1-t) P_1 + t^2 P_2 \]  

(5)

while cubic Bezler curves have four control points and are defined by

\[ B(t) = (1-t)^3 P_o + 3 (1-t)^2 t P_1 + 3 (1-t) t^2 P_2 + t^3 P_3 \]  

(6)

These Bezler parametric curves are extensively used in the design of highways and railways, CAD systems, design of prototypes in many industrial applications[17,18], mesh
design in FE environments\cite{19}, and more recently in the geometrical approximation of fibers in fiber reinforced composites\cite{20}.

In what respect to failure and yielding surfaces, it can be mentioned the work by Vegter et. al\cite{21} who proposed a representation of the yield function using Bezier polynomials, interpolating mechanical tests results directly for the case of anisotropic metallic sheets, under a biaxial stresses state.

These curves have not yet been used in the field of mechanics of cohesive frictional materials. In this paper both quadratic and cubic Bezier curves are considered for the interpolation function.

4 INTERPOLATION BETWEEN COMPRRESSIVE AND TENSILE MERIDIANs BASED ON QUADRATIC BEZIER CURVES

In this section a quadratic Bezier interpolation function is proposed. In the deviatoric partial view in Fig. 3 the location of the three control points, as well as the adopted cartesian coordinate system can be observed.

\begin{equation}
\begin{aligned}
P_0 &: x_0 = 0; y_0 = \rho_t; \rho_o = \rho_t \\
P_1 &: x_1 = \frac{\sqrt{3}}{3} (2\rho_c - \rho_t); y_1 = \rho_t; \\
P_2 &: x_2 = \rho_c \cos \left( \frac{\pi}{6} \right) = \frac{\sqrt{3}}{2} \rho_c; y_2 = \rho_c \sin \left( \frac{\pi}{6} \right) = \frac{1}{2} \rho_c; \rho_2 = \rho_c.
\end{aligned}
\end{equation}

The parametric expressions result
Nevertheless, in this case, contrary to the case of CAD applications where it is not important to relate parameter \( t \) with an angle, here it results crucial to define the relation \( t - \theta \). This can be derived from the expression of the tangent of the Lode angle \( \tan\theta = \frac{x}{y} \), leading to a quadratic equation which solution is

\[
t = \frac{2(e - 2) + \sqrt{4(2 - e)^2 + \{4e - 5 + \sqrt{3(2e - 1)}\tan\theta\}}}{\{4e - 5 + \sqrt{3(2e - 1)}\tan\theta\}} \tag{10}
\]

Finally, factor \( r \) in Eq. 2 can be obtained by the following equation

\[
r = \frac{\rho_c}{\rho} = \frac{\rho_c}{y} \cos \theta = \frac{2 \cos \theta}{(2e + (1 - 2e)t^2)} \tag{11}
\]

If Eq. 10 is replaced in Eq. 11, then a similar expression to that in Eq. 3 is obtained, leading almost to the same curve as the elliptical interpolation, with the disadvantage that presents a singular point for \( \theta = 30^\circ \) and \( e = 0.80 \). Thus, the only goal of replacing the elliptical interpolation by quadratic Bezier curves is the parametric definition of the curve.

5 INTERPOLATION BETWEEN COMPRESSIVE AND TENSILE MERIDIANS BASED ON CUBIC BEZIER CURVES

In this section a cubic Bezier interpolation function is proposed. In the deviatoric partial view in Fig. 4 the location of the four adopted control points.

In this case, continuity conditions at \( \theta = 0^\circ \) and \( 60^\circ \) are not enough to define the geometry of the control polygon. For defining that geometry, the concept of a shape control parameter was taken from Habib and Sakaib (2009)\textsuperscript{[18]}, where cubic Bezier functions are used to trace the transition of two circles. It should be said that this case is completely different to the one presented in that work, and therefore the formulation is different.

The hypothesis to define the intermediate two points \( P_1 \) and \( P_2 \) was to trace a line parallel to the one joining \( P_o \) and \( P_4 \), which forms an angle \( \gamma \) with respect to an horizontal line. Then a shape parameter \( \alpha^h \) was introduced to define the distance between \( P_o \) and \( P_1 \). It can be demonstrated that it must be \( 1/3 < \alpha^h < 2/3 \), but the optimum values, adopted in this paper are \( 0.40 \leq \alpha^h \leq 0.60 \). Under this hypothesis the control points result

\[
x(t) = \frac{\sqrt{3}}{6} \rho_c \left( 4(2 - e)t + (4e - 5)t^2 \right) \tag{8}
\]

\[
y(t) = \frac{1}{2} \rho_c \left( 2e + (1 - 2e)t^2 \right) \tag{9}
\]
Figure 4: Control Points in the adopted cubic approximation using Bezier curves

\[ P_0 : x_0 = 0; \quad y_0 = \varepsilon \rho_c \]
\[ P_1 : x_1 = \left(1 - \alpha^h\right) \left(2 - e\right) \frac{\sqrt{3}}{3} \rho_c; \quad y_1 = \varepsilon \rho_c \]
\[ P_2 : x_2 = \left(\alpha^h \left(2e - 1\right) + 2 \left(2 - e\right)\right) \frac{\sqrt{3}}{6} \rho_c; \quad y_2 = \left(2e \left(1 - \alpha^h\right) + \alpha^h\right) \frac{1}{2} \rho_c \]
\[ P_3 : x_3 = \frac{\sqrt{3}}{2} \rho_c; \quad y_3 = \frac{1}{2} \rho_c \]

Figure 5: Deviatoric predicted shapes - Comparison between the elliptical and the cubic Bezier approximations, considering two different values of parameter \(\alpha^h\) and an eccentricity \(e = 0.55\)
The parametric expressions result

\[
x(t) = \left\{ (1 - 3\alpha_h) t^3 + (3\alpha_h + (1 - \alpha_h) (2e - 4)) t^2 + 2 (1 - \alpha_h) (2 - e) t \right\} \frac{\sqrt{3}}{2} \rho_c \quad (13)
\]

\[
y(t) = (2e - ( (1 - 3\alpha_h) t^3 + 3\alpha_h t^2) (2e - 1)) \frac{1}{2} \rho_c \quad (14)
\]

In the cubic case, the relation \( t - \theta \) must be obtained from a cubic equation. This equation was developed and the obtained result was used to trace the plots in Figs. 5 and 6, for different values of \( \alpha_h \) and for two values of eccentricity, \( e = 0.55 \) and \( e = 0.80 \), respectively. Using cubic Bezier curves, no singularity point is observed and moreover, parameter \( 0.40 \leq \alpha_h \leq 0.60 \) permits to modify the shape of the deviatoric view.

The main goal of this proposal can be observed by analyzing the predictions in the biaxial stresses space. For this purpose, the compressive and tensile meridians were defined based on the Performance Dependent Failure Criterion (PDFC) \(^{[11]}\), and then the proposed cubic interpolation function was applied.

In Fig. 7, a comparison between the predicted peak stresses based on the cubic Bezier interpolation for \( \alpha_h = 0.60 \) with experimental results obtained from Hussein and Marzouk (2000) \(^{[22]}\) is presented. In that figure, the predicted strengths based on the elliptical interpolation can also be seen represented by the dashed line. This comparison based on the elliptical interpolation was previously performed (See \(^{[13]}\)) and a relative error of 8.40% was reported. It can be observed in Fig. 7 that using the proposed cubic Bezier interpolation that error reduces to only 2.40%.
6 CONCLUSIONS

- In this paper the use of quadratic and cubic Bezier curves was explored for the interpolation between the compressive and tensile meridians of the failure surface of concrete like materials in order to define the deviatoric view of the surface. The proposed functions are also valid for yielding surfaces during hardening or softening processes.

- It was showed that a quadratic approximation does not present important advantages in comparison with the classical elliptic interpolation by Willam and Warnke (1974)\cite{8}.

- On the contrary, it was demonstrated that a cubic Bezier approximation involving the definition of a shape parameter, compared against the elliptic interpolation, can considerably improve the accuracy of the prediction of peak stresses for intermediate Lode angles, particularly in the case of biaxial stress states, without loosing the benefits of continuity. Nevertheless the computational cost is higher for the cubic Bezier approximation.

- In the future, the benefits of using the proposed cubic Bezier approximation leading to a parametric expression of the deviatoric shape of failure and yielding surfaces should be explored.
7 Acknowledgments

The authors gratefully acknowledge the partial financial support of this work by the University of Buenos Aires (Project UBACyT 2014-2017 20020130100769BA). They also acknowledge the support provided by the Institute of Science Technology and Engineering "Hilario Fernández Long" (INTECIN) and by the Laboratory of Materials and Structures at FIUBA, Argentina.

References


Paula Folino and Damián Smilovich


Key words: Pendulum, Variable length, Swing up, Simulink, Inertia.

Abstract. This paper presents the study of a system composed by a simple pendulum with distributed mass carrying a control device that is able to adjust its length. This will result in an extra injection (or loss) of energy in the system, according to the control strategy. The system is modeled as a body with variable inertia, and for computations, Simulink software, from MatLab, is used. Instantaneous change will be compared to continuous variation in length. In a second phase, friction is added to the system, and a way to compensate this friction using the change of the length is proposed. To view more clearly this problem, it is possible to make an analogy with a person sitting on a swing. The person can move his legs in order to leave them in two ways: straight or bent. This way, the system swing-person may be seen as an auto parametrically excited system, due to the variation in the position of the center of mass. This article focuses on a broader topic when control is used to alter the distribution of mass of a mechanical system. The objective of this work is to develop a control technique to make the pendulum do a complete revolution only by varying its length and then decelerate till it stops vertically.

1 INTRODUCTION

An important tool in modern mechanical engineering is the possibility of visualization of the motion of a considered system and, if we have a mathematical model, the possibility of calculating forces and optimizing parameters. So even in classical problems sometimes it is possible to have new insights. Since Galileo’s study on motion of pendulums in the beginning of the 17th century, many pendulum-like systems have been studied and built, and the knowledge on their dynamics seems to be complete. But this kind of simple system may represent well a more complex body combinations and it is always possible to find something that is new.

The behavior of pendulum systems is more or less influenced by disturbances, nonlinearities, friction and faults. This article presents a variation of the simple pendulum, with a device that extends its size. In the conventional problem, a simple pendulum can’t swing back to a higher angle and also, if there is added friction, the pendulum dissipates energy till it stops in a vertical position. In the present case, adjusting its length during a swing will add or subtract energy from the system.
This system was thought by observing people swinging in a park. In order to swing properly, one must adjust their legs and body so the center of mass of the system swings-person changes and it creates a variation on the system’s inertia. Small children normally needed a parent’s assist to help swing back and forward, due the lack of leg size or coordination. Grownups and bigger children do not have this difficulty and can swing by themselves. In addition, if we had a rigid chain on the park swing, maybe any person could do a loop, despite the danger.

Our system consists on a rigid hinged hollow bar with distributed mass, with a mass concentrated on the edge of the fixed part, simulating an actuator that operates a movable part which we call rod and may retract or extend from the hollow bar. This movable part is controlled according to the desired strategy, increasing the energy of the rod or decreasing it. The final aim of the study is to complete a full rotation of the pendulum only by extending the rod.

For such study, classic models of inverted pendulum were used (Åström & Furuta, 2000) and control strategies of swing up are discussed (Bugeja, 2003) (Chung and Hausser, 1995) (Magnus, 1967), also with strategies with a dissipative term (Brogliato, Lozan and Maschke, 2007). An approach of switching control (Zhai, 2007) is presented to case the movement after one revolution of the pendulum using friction or not. Finally, the model is simulated with the software Simulink, by using default control libraries (Jadlovská and Sarnovský, 2011) and real world values and in the end the results are shown.

2 MODELING

2.1 System model

The model used for the study consists on a bar L1 that may rotate freely around a pivot, and an extension L2 that varies during time, therefore:

\[ L1 + L2 = L(\alpha, \dot{\alpha}) \] (1)

An angle \( \alpha \) is measured from the vertical. The pendulum has an initial condition of \( \alpha \) different from zero and \( \pi \). Considering the angular momentum of the pendulum with reference on the fixed point \( O \) is:

\[ \mathcal{L} = mL^2\dot{\alpha} \] (2)

In addition, the moment of the force of gravity is:

\[ \mathcal{M} = -mgL \sin \alpha \] (3)

Derivating equation (2) and equaling to equation (3):

\[ \frac{d}{dt}(mL^2\dot{\alpha}) = -mgL \sin \alpha \]

\[ 2mL\ddot{\alpha} + mL^2\ddot{\alpha} = -mgL \sin \alpha \]
\[
\ddot{\alpha} + \frac{2\dot{L}}{L} \dot{\alpha} + \frac{g}{L} \sin \alpha = 0
\]  

(4)

In the regular pendulum, the \( \dot{\alpha} \) component is equal to zero due \( L \) being constant, which gives the classic pendulum equation. This component also makes it an equation for parametrically excited oscillators.

![Figure 1: Model of the self-excited pendulum](image)

Also, in the case the friction is added to the system, the new equation is:

\[
\ddot{\alpha} + \left( \frac{2\dot{L}}{L} + b \right) \dot{\alpha} + \frac{g}{L} \sin \alpha = 0
\]  

(5)

Where \( b \) is the constant for viscous friction.

2.2 Control technique

All referenced articles and books use a nonlinear control technique to swing up the pendulum. Specifically Magnus uses an open-loop nonlinear control of \( L(t) \) using sign of the angle \( \alpha \).

\[
L(\alpha, \dot{\alpha}) = L1 + \frac{L2}{2} (1 - sgn(\alpha) sgn(\dot{\alpha}))
\]  

(6)

The result of that is a Boolean control that changes the extension on the rod to a maximum or a minimum length, no intermediate value. Of course, that instant change makes the system
only theoretical, since the necessary infinite acceleration and deceleration of the extended part will not be reach in the real world.

![Figure 2: Boolean control used as reference for the study](image)

To make the numerical study clothier to reality, we opt to change the function sign for the actual angle. Then, the equation becomes:

\[
L(\alpha, \dot{\alpha}) = L_1 + \frac{L_2}{2}(1 - a\dot{\alpha})
\]  

(7)

In this way the derivate of \(L(\alpha, \dot{\alpha})\) is no longer infinite, and when the pendulum reaches its maximum value on the swing, the transition from one length to the other is not instantaneous:

![Figure 3: System controlled by the suggested equation](image)

It’s possible to notice that the component \(-a\dot{\alpha}\) is responsible for the acceleration on the pendulum. Therefore, to slowdown the pendulum after one revolution the control should switch to the following equation:

\[
L(\alpha, \dot{\alpha}) = L_1 + \frac{L_2}{2}(1 + a\dot{\alpha})
\]  

(8)
With this change, the system will behave like the one shown in figure 3, except that the arrows will be in the opposite direction.

A switching control is responsible for changing the strategy, deciding how many loops the system will do before decelerating. This strategy then makes the extension of the rod retract or enlarge, changing the inertia of the system by injecting or removing energy in the system.

![Control schematic](image)

**Figure 4:** Control schematic

3 SIMULATIONS

The system was modeled in MATLAB’s Simulink. We are interested in surpass the damping effect of the viscous friction, and accelerate the pendulum and decelerate once it completes a revolution. For such, we developed the system in the figure 6.

The model was built according to the equation (4). For L were used equation (6) for the acceleration and equation (7) for deceleration. Both equations are operated by a simple logic control that checks if the angle is less than π. If true, the pendulum will continue to accelerate, enabling only the Acceleration box. If false, it means the edge of the pendulum has reached the maximum height and it begins to descend, by enabling only the Deceleration box.
Since it’s important to assume that the system is plausible in the real world, a limit to the operation of the extend actuator velocity is added, and the value is 3 m/s. This value is seen in the literature as a very fast actuator, considering no weight on the edge of the piston.

Some values used for this simulation are:
Table 1: Simulation constants

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>0.5 m</td>
</tr>
<tr>
<td>L2</td>
<td>0.2 m</td>
</tr>
<tr>
<td>b</td>
<td>0.2 Nms</td>
</tr>
<tr>
<td>$\alpha_0$</td>
<td>$\pi/8$ rad</td>
</tr>
<tr>
<td>dL limits</td>
<td>$\pm 3$ m/s</td>
</tr>
<tr>
<td>$t_{simulation}$</td>
<td>20 s</td>
</tr>
</tbody>
</table>

As we can see on figure 7, it was possible to inject energy into the pendulum only by changing its size, which caused an increase in the oscillation of the system, and when it reached the highest spot (marked as a red dot on the graphs) the deceleration of the pendulum lead to a damping in the movement of the rod. This controlled damping is also bigger than a normal viscous damping.

Furthermore, varying the friction to a lower value it’s possible to see that the maximum velocity of the extend actuator (dL) diminishes during the acceleration step. This happens because the swing of the pendulum has less energy loss comparing with the last simulation when the friction is higher. The energy added to achieve the lap in this case is lower, which means less velocity needed. This result has a big impact on the direction of the movement, making the system turn counter-clockwisely, unlike the other simulation.
The setback in this case is the deceleration taking more time to reach a motionless state.

![Figure 8: Simulation outputs with \( b=0.02 \)](image)

4 CONCLUSIONS

We presented a well-known system studied in a different way. By adding a mechanism capable of changing the inertia of the system, it was possible to move a pendulum without adding an actuator to the fixed point, or adding external forces to the system. If we think about real embedded unbalanced systems, for example, where it is not possible to add a control device in a certain way, alternatives should be studied to ensure a perfect operation.

The modeled system worked as expected, though unexpected results about the real effect of viscous friction opened our eyes. By reducing this friction the clockwise loop turned into a counter clockwise loop. The effect of the maximum velocity of the extend actuator is also crucial to the success, since there is a relation between friction and actuator speed that makes it work or not.

The open-loop controller followed a simple strategy of accelerating and decelerating the pendulum. This makes it a solution for the problem, but not the only. Another solution could use a simple closed-loop controller, or an adaptive controller.

Every result opens a couple of doors to new ideas using inertia variation to control systems. These ideas bring solutions to problems seen in everyday life, or even to problems created while sitting on a park bench.
5 ACKNOWLEDGMENT
The authors acknowledges helpful discussions with our colleagues from Dynamics and Vibrations Laboratory. Support for this work was provided through the Mechanical Engineering Department of Pontifical Catholic University (PUC-Rio) and CAPES.

REFERENCES
SIMULATION OF AN ACTIVE SUSPENSION USING PID CONTROL

SOUSA D.F.¹ AND AVILA S.M.²

¹ UnB Engineering Faculty Gama, University of Brasilia
Área Especial de Indústria Projeção A Setor Leste Gama/DF CEP: 72.444-240
darlanferreira.s@hotmail.com www.fga.unb.br

² UnB Engineering Faculty Gama, University of Brasilia
Área Especial de Indústria Projeção A Setor Leste Gama/DF CEP: 72.444-240
avilas@unb.br www.fga.unb.br

Key words: vehicle dynamics, active suspension, PID control.

Abstract. The suspension is the automotive system that connects the vehicle sprung mass to the unsprung mass relative to wheels and tires. Its main proposal is to improve passenger comfort and increase safety in operation, providing a better tire-road contact. A vehicle suspension can be classified as passive, active or semi-active. Passive suspensions have come a long time ago being widely used on automotive industry. The properties of mechanical components cannot be modified in real time via external signals, the vibration is controlled only storing or dissipating the energy associated with the movement through its main components such as springs, dampers and masses. The principle of an active suspension is to install an actuator between the sprung and unsprung masses in addition to a spring and a damper, or replacing both of them. Sensors are placed at different points of the vehicle to measure the vibrations. The information gathered by sensors is sent to a controller that calculates in real time the actuator force magnitude that must be applied. In semi-active system suspensions, damping properties and stiffness can be changed in real time by a control signal. In this work a quarter car suspension model is simulated in order to study an active suspension with an automatic PID controller. This automatic control can be divided in proportional, integral and derivative controls or alternatively a combination between them. Results are compared to passive suspension considering different road surfaces. A good performance is achieved with the PID active suspension reducing vibration amplitudes what improves ride comfort.

1 INTRODUCTION

A vehicle suspension system is a mechanism composed by springs, shock absorbers and linkages that connect the vehicle body to wheels, reducing the vibration generated and felt by the passengers, when the vehicle is operating in a given field [1]. The main challenge of an automotive suspension design is suit stability and passengers/driver comfort. It has to be soft enough to ensure human comfort, reducing car body acceleration due to uneven ground, and hard enough to guarantee solid contact of the wheel to the ground, increasing stability.

Passive suspension has been widely used in automotive industry for decades. In this kind
of system mechanical components properties cannot be modified in real time, the response is controlled due to energy storage or dissipation, associated with its main components motion, such as springs, dampers and masses [2]. Due to its low cost and easiness, passive suspension still remain competitive and are widely used, being the type most commonly used in the Brazilian industry.

However passive suspensions have the disadvantage of not responding well to uneven ground [3]. The development of new sensors and actuators, control algorithms, modelling techniques as well as computational tools enable active suspension evolution.

Active suspension principle is to insert an actuator between sprung mass and unsprung mass besides spring and damper, or replacing them. Sensors are placed at different points in the vehicle to measure its motion, besides the suspension and unsprung motion also. The information collected by sensors is sent to a real time controller that calculates the force magnitude that has to be applied by the actuator [4]. An automatic controller compares the output real value with the reference value (target value), determines the deviation and produces a control signal which will reduce the deviation to zero or a small value [5].

Florin et al [6] simulated a vehicle suspension with a quarter car model and studied its performance subjected to a step excitation. The mathematical model describing the system was both in second order differential equations as state space. Numerical analysis was carried out in MATLAB computing environment (MathWorks), using Simulink toolbox.

Agharkakli [1] proposed a mathematical model for passive and active suspension. And in addition implemented the technique of Linear Quadratic Control (LQR) in a model of a quarter car vehicle for active suspensions, its performance was compared with the passive suspension submitted to different terrain profiles.

Cavalheiro and Avila [7] performed parametric studies on passive suspension with quarter and half car models, analysing system performance with different inputs. Thus, it was found that the performance of passive suspension depends largely on the type of land on which it operates. Certain parameter values can be satisfactory on a particular ground, not being efficient in another.

Natsiavas et al [8] proposed a methodology to optimize suspension parameter in cases of random ground excitations. First a passive suspension system was considered then an approximate active system with sky-hook damping was analysed.

Changizi e Rouhani [9] implemented a fuzzy control technique to a continue damper suspension and compared its performance with PID (Proportional integral derivative) control technique to different road profiles via numerical simulation.

Fayyad [10] designed a PID controller including hydraulic dynamics for a vehicle quarter car model, using software MATLAB, modelling a PID controller, to improve the ride comfort and road holding ability. It was found that active suspension system improves ride comfort even at resonant frequency.

This work studies the performance of an active suspension, with a quarter car vehicle model using a PID controller. Controllability and observability properties are analysed in a way to verify good control performance. The analysis is carried out using MATLAB and SIMULINK toolbox capabilities. It is found out that with optimized proper PID parameter values, a good performance is verified for different types of road profiles.
2 MATHEMATICAL FORMULATION

Suspension systems can be modeled approximately by spring-mass-dashpot system of two degrees of freedom, as the one shown on Fig. 1. It is called a quarter car model, where the sprung mass $M_s$ is attached by the suspension, modeled as a spring and a damper, to the unsprung mass $M_u$. The spring stiffness is given by $k_s$ and the damping coefficient is $c_s$. When considering an active suspension, the automatic actuator is connected and is modeled by a control force $u$. The tire is represented by a spring with stiffness $k_t$, not considering damping due to tire viscoelastic properties. This model is restricted only to vertical vehicle dynamics analysis.

Figure 1: Quarter car model [5]

According to Newton’s Second Law the motion equation system can be obtained:

$$m_u \ddot{x}_2 = k_s (x_1 - x_2) + b_s (\dot{x}_1 - \dot{x}_2) - k_t (x_2 - w) \tag{1}$$

$$m_s \ddot{x}_1 = -k_s (x_1 - x_2) - b_s (x_1 - x_2) \tag{2}$$

Where $x_1$ is the sprung mass displacement, $x_2$ is the unsprung mass displacement and $w$ is the dynamic excitation due to the road profile. Equation (3) shows the motion equation on a matrix form:

$$\begin{bmatrix}
m_s & 0 \\
m_u & m_l
\end{bmatrix}
\begin{bmatrix}
\dot{x}_1 \\
\dot{x}_2
\end{bmatrix}
+ \begin{bmatrix}
k_s & -k_s \\
-k_s & (k_t + k_s)
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix}
+ \begin{bmatrix}
bs & -bs \\
-bs & bs
\end{bmatrix}
\begin{bmatrix}
\dot{x}_1 \\
\dot{x}_2
\end{bmatrix}
= \begin{bmatrix}
0 \\
(k_t \times w)
\end{bmatrix} \tag{3}$$

The transfer function $G(s)$ associated to the output, corresponding to the sprung mass displacement, due to the road profile input is given by:

$$G(s) = \frac{k_t (bs + k_s)}{m_u m_s s^4 + (m_s + m_u) bs^3 + [k_t m_s + (m_u + m_s) k_s] s^2 + k_t bs + k_t k_s} \tag{4}$$
The motion governing equation can also be rewritten in a state space form:

\[
\dot{Z} = Az + Bu
\]

\[
Y = Cz + Du
\]

Where \( A \) is the state space matrix, \( B \) is the input matrix, \( C \) is the output matrix, \( D \) is the direct transmission matrix and \( U \) is the input of system. To the quarter car system presented in Fig. 1 matrices \( A \) and \( B \) are given by:

\[
A = \begin{bmatrix}
0 & 0 & 1 & 0 \\
0 & \frac{ks}{ms} & \frac{bs}{ms} & \frac{bs}{mu} \\
\frac{ks}{ms} & \frac{ks}{ms} & \frac{bs}{ms} & \frac{bs}{mu} \\
\frac{ks}{mu} & \frac{kt + ks}{mu} & \frac{bs}{mu} & \frac{bs}{mu}
\end{bmatrix} ;
B = \begin{bmatrix}
0 \\
0 \\
0 \\
\frac{kt}{mu}
\end{bmatrix}
\]

The output matrix \( C \) can be determined according to the degrees-of-freedom output that can be measured from the system. In this work, in order to calculate the observability of the system in different settings, matrix \( C \) will be varied to verify the influence on observability property. The matrix \( D \) of direct transmission will not be considered in this work and is set to zero, a null matrix.

A system is called a controllable system at an initial time if it is possible through an adequate control vector transfer the system from any initial state to any other state, in a finite time interval. A system is considered observable if in an initial time it is possible to define its state through the output observation during a finite time [5]. These concepts are important because they may provide a solution to the control design. A control system cannot be a solution if the system is uncontrollable. So knowing these parameters, it is necessary to facilitate the resolution of the desired control system.

PID control (proportional-integral-derivative) compares the real value of the output quantity with the reference value (target value), determines the deviation and produces a control signal which will reduce the deviation to zero or a small value [5]. The transfer function in this case is given by

\[
\frac{P(s)}{E(s)} = K_p (1 + \frac{1}{T_i s} + T_d s)
\]

Where \( K_p \) represents the proportional gain, \( T_d \) represents the time derivative and \( T_i \) the integral time. The first term of the transfer function corresponds to proportional control, integral control to the second and so on. The Simulink® software tool MATLAB®, has the simulation of PID controller in the form of block diagram. In the controller configuration there is a Tune tool that enables tune up the PID controller gains automatically, it calculates the linear model of the plant, combining all the input and output controller blocks, including all control loop blocks.
The algorithm used in automatic PID, meets the robustness and performance goals through an initial controller design, choosing a bandwidth to achieve project equilibrium, based on open-loop frequency response of the linear model. When changing the response time, controller bandwidth or phase margin, the algorithm calculates a new gain [11].

The open loop block diagram used to analyse the passive suspension is show on Fig. 3. In active suspension case the block diagram is a closed-loop diagram shown in Fig. 4. A set point is selected to a constant value, in this work a constant value of zero was defined [10].

![Open-loop block diagram](image1)

![Closed-loop block diagram](image2)

3 NUMERICAL RESULTS

Freitas [12] suggested the properties of the quarter car suspension system from an approximation of a quarter car model to a real MacPherson suspension. Mass and damping parameters were determined assuming it is a popular vehicle according to design suggestions of Gillespie [13]. The parameters are specified in Tab. 1.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sprung mass (Kg)</td>
<td>Ms</td>
<td>203,0</td>
</tr>
<tr>
<td>Unsprung mass (Kg)</td>
<td>Mu</td>
<td>26,0</td>
</tr>
<tr>
<td>Spring stiffness (N/m)</td>
<td>Ks</td>
<td>23,000,0</td>
</tr>
<tr>
<td>Tire stiffness (N/m)</td>
<td>Ki</td>
<td>200,000,0</td>
</tr>
<tr>
<td>Damping coefficient (Ns.m)</td>
<td>Bs</td>
<td>1,200,0</td>
</tr>
</tbody>
</table>

In a way to verify system controllability and observability, Matlab capabilities were used. Giving A (State space matrix), B (Input matrix) e C (Output matrix) it is possible to calculate controllability and observability matrices through ctrb and obsv commands, respectively.
Then the rank of these matrices is calculated. The controllability matrix rank was 4 indicating that the system is controllable. In the case of observability property, three different cases were studied: a) measuring only $x_1$; b) measuring only $x_2$; c) measuring both $x_1$ and $x_2$. In all cases the rank of the observability matrix was 4 indicating that the system is observable no matter measuring only one output or both of them.

In sequence, the suspension system was subjected to three different road profiles: step; harmonic and white noise, shown in Figs. 3, 4 and 5.

![Figure 3: Step function of 0.1 m](image)

![Figure 4: Harmonic profile $W = W_0 \sin(10.07t)$](image)
Figure 5: Random profile – White noise

Figure 6 shows sprung mass time history when the passive quarter car system is subjected to the step road profile. It can be observed an overshoot of 53% and a suspension time response of 3.1 s. These values can be minimized to improve passengers comfort by the installation of an active suspension.

Figure 6: Sprung mass displacement time history when subjected to a step load profile

Designing a PID controller, as explained previously, is proposed to an active suspension. Figure 7 shows both responses: sprung mass displacement for passive and active suspensions to a step road profile. Active suspension improve the system behavior with an overshoot of only 29%, reducing the maximum displacement on 45.3%. Suspension time response reduced from 3.1 s to 2.3 s (25.8%).
Figure 7: Sprung mass displacement time history – step road profile

Figure 8 shows both responses: sprung mass displacement for passive and active suspensions to a harmonic road profile. It can be observed that the maximum displacement reduced about 83.8% comparing active to passive case. And also on the steady state response a good improvement on performance is achieved.

Observing Figure 9, that shows both responses: sprung mass displacement for passive and active suspensions to a white noise road profile, it can be noticed that also in this case active PID suspension achieves a very good performance. When comparing maximum sprung mass displacement a 72.5% reduction is reached.
Figure 8: Sprung mass displacement time history – harmonic road profile

Figure 9: Sprung mass displacement time history – White noise road profile
4 CONCLUSION

This study performs numerical simulations of a quarter car model in order to check the vertical dynamic response, when the system is excited by an irregular road profile. Three types of excitation were considered: step, harmonic and white noise. The purpose was to evaluate the behavior of an automatic feedback control improving comfort and handling.

The system showed up to be controllable and observable. Comparing the performance of an active suspension, designed with a PID controller, with the passive one, it can be observed a considerable improvement on efficiency for all the road profiles considered. In the case of the harmonic road profile a reduction of 83.8% on the maximum sprung mass displacement was found out.

Setting the PID parameters through Tune Simulink tool, mentioned in the article, is merely a basis for designing the controller, it is recommended verify what is the influence of the PID gains on the behavior of the controller to make a more detailed analysis.

It has been found that the action of PID control showed satisfactory results improving the performance of the suspension system, however active suspensions still have a very high cost of manufacturing, installation and maintenance compared to passive suspensions, that is the reason it is not widespread in series productions of vehicles.

ACKNOWLEDGEMENTS

The authors acknowledge the financial support for this research to CNPq (National Council for Scientific and Technological Development).

REFERENCES


SYNTHESIS OF MULTI-FINGERED ROBOTIC HANDS

Alba Perez-Gracia

1Department of Mechanical Engineering
Idaho State University
Pocatello, ID 83202, USA
perealba@isu.edu

Key words: Kinematic Synthesis, Multi-fingered hands

Abstract. Wristed, multi-fingered hands have recently been the target of rigid body
guidance dimensional synthesis, for simultaneous tasks of all the fingertips. Solvability
conditions have been derived for general hand topologies, consisting on common joints
and several branching stages, to end in an arbitrary number of end effectors, or finger-
tips. Synthesis equations have been created and solved numerically, to obtain new and
innovative hand designs. In the recently developed design tool, positions, velocities and
accelerations of the fingertips can be defined.

Here we present the latest developments in the systematic design of wristed, multi-
fingered hands, including input data analysis, synthesis equations, solvability and appli-
cations to new families of hands. Some examples are included.

1 INTRODUCTION

Kinematic chains with a tree topology consist of several common joints that branch to
a number of serial chains, each of them corresponding to a different end-effector. A typical
example of a kinematic chain with a tree topology is a wristed, multi-fingered hand.

Compared to other topologies, the tree topologies have not been so widely studied.
Kinematic analysis for applications in modular robots and robotic hands can be found
in [16] and [18], and dynamic analysis is found in [3] and [5]. Structural synthesis for
multiple fingers with no wrist, considering grasping and manipulation requirements, are
found in [7]. The first reference to kinematic synthesis of tree topologies is found in [12].

The kinematic synthesis of these topologies presents particular challenges that are
different of those that appear in single serial chains or in closed-loop systems. In particular,
the kinematic synthesis of multi-fingered hands has been explored also in [14] and more
extensively in [13]. In these works, the foundation for the kinematic synthesis were laid.

Here we combine all the current results in task definition, type synthesis and dimen-
sional synthesis, in order to present a complete design methodology for multi-fingered
robotic hands. The goal is to provide robotic designers with tools to help in the design of more efficient and better designed robotic hands.

2 TOPOLOGY OF ROBOTIC HANDS

A tree topology for a kinematic chain has a set of common joints splitting on several chains, ending in multiple end-effectors [11]. The tree topology is represented as a rooted tree graph, where each vertex can be connected to several edges. The approach of Tsai [19] is followed, with the root vertex being fixed with respect to a reference system.

Wristed, multi-fingered hands are kinematic chains with a tree or hybrid topology. Figure 2 shows some well-known hands and their graph representation.

![Figure 1: Some hands and their graph representation.](image1)

2.1 Compacted and Reduced Graphs

When formulating a synthesis problem, the internal loops in the hand structure can be substituted using a reduction process [13], so that the hand has a tree topology with links that are ternary or above. The reduction eliminates the loops and substitutes those with the most constrained path of that loop. Figure 2 shows the graph representation of a hand that has an internal loop.

![Figure 2: Original graph, left; compact graph, center; reduced graph, right.](image2)
2.2 Tree Topologies

Tree topologies are denoted as $W - (B_1, B_2, \ldots, B_b)$, where $W$ are the common joints and the dash indicates a branching or splitting stage, with the branches contained in the parenthesis, each branch $B_i$ characterized by its type and number of joints. In the case of using just revolute joints, the joint type is dropped and only the number of joints is indicated. Figure 3 shows the compacted graph for the $R - (R, R, R, R, R)$, or $1 - (1,1,1,1,1)$ chain. This is one of the member of the one-jointed hand family, in particular the five-fingered hand. The root vertex is indicated with a double circle.

![Figure 3: Tree graph for the 1 – (1,1,1,1,1) robotic hand with one palm, left; fractal hand with two splitting stages and three branches per split, right.](image)

A hand topology is represented by two arrays. We assume a numbering of the graph edges and define a *parent-pointer array* and a *joint array*. The length of both arrays is equal to number of edges of the graph after the reduction process is applied.

![Figure 4: A five-fingered, two-palm hand topology. (a) indicates the numbering of the edges and (b) indicates the number of joints for each edge.](image)

The parent-pointer array implements the parent-pointer representation, where each element takes the value of the previous edge, the first edge being the one incident at the root vertex. Each element of the joint array contains the number and type of joints for each edge. As an example, for the tree topology shown in Figure 4, the parent-pointer array and joint array are defined as $p = \{0,1,1,1,3,3,3\}$ and $j = \{2,2,1,2,3,3,3\}$. 
2.3 Some Families of Tree Topologies

Compacted and reduced tree topologies can be grouped in families that share some common characteristics. The family of robotic hands with one branching and identical fingers was studied in [9], and the particular case of one-jointed wrist and fingers has been synthetized in [10]. The family of binary hands, one of the subfamilies of the family of fractal hands, has been studied in [4].

- Hands with Identical Fingers: Hands with a wrist consisting of \( p \) revolute joints and a single palm spanning \( b \) branches or fingers, each of them having \( q \) revolute joints. This is the \( p - (q, \ldots, q) \) family of hands. Some of the members of this family are solvable regardless of the number of fingers. The graphs of this family are all quite similar, Figure 3, left, being one example of a tree from this family.

- Fractal hands: These hands have a number \( s \) of splits, each of them splitting in the same number of branches, \( b \). The number of end-effectors of this family grows quite quickly; even for the smaller subfamily, the binary hands, the number of end-effectors after \( s \) splits is \( 2^s \). An example of a fractal hand is the \( 1 - (1 - (1,1,1), 1 - (1,1,1), 1 - (1,1,1)) \) hand presented in Figure 3, right, with three branches per split.

3 KINEMATIC TASKS FOR ROBOTIC HANDS

A kinematic task is usually defined as a set finitely-separated positions, such as the ones shown in Figure 5. Velocities, or even accelerations, can be defined at some of those positions.

![Figure 5: A kinematic task defined by a set of positions.](image)

Kinematic tasks for robotic hands need to specify the position of each fingertip or contact surface; we denote this as a simultaneous task.

3.1 Motion Capture for Kinematic Tasks

The difficulty of defining three-dimensional motion tasks for rigid bodies is well known [6]; as a consequence, the data is usually generated by sensing the desired motion, as
performed by a moving element. Motion sensing methods provide with a realistic and quick set of data points; however, noise from sensors and other, less known problems need to be addressed. One of them is the excessive richness of the data, especially when it is human-generated. In our experience, small motions in the many degrees of freedom of the human body appear even when the subject is trying to perform a single, 1dof motion \cite{20}. Those mini-motions need to be identified the dataset. Another issue is the dealing with excess of information, commonly with hundreds of data points per seconds. Using all those points leads to a costly optimization process, while in downsampling we may be missing key parts of the motion while keeping uninteresting segments.

For synthesis purposes, a reference frame is tracked in order to obtain rigid-body motion. Figure 6 shows the experimental setup used for the motion capture with reflective markers and infrared cameras. Recent work \cite{8} uses principal component analysis to select the significant motion from the data, and data association techniques to extract a desired number of significant positions from the dataset.

![Figure 6: Experimental setup for motion capture.](image)

### 3.2 Lie Groups and Lie Algebras

While the spatial displacement approach allows to minimize the number of points to track for a rigid body, the application of the above-mentioned techniques presents new challenges in the fact that the rigid body displacements, as a Lie group, do not form a Euclidean manifold.

In order to avoid this problem, a common technique \cite{15} is to work on the Lie algebra of the Lie group, by using the logarithmic map. The physical interpretation of this is to consider that the finite motion is performed along the shortest, or straight path, of a screw motion about a fixed axis.

Let $\hat{q}$ be an element of the Lie group $SE(3)$, a finite displacement expressed for instance as a dual quaternion. Consider the corresponding element of the Lie algebra $w \in se(3)$ as being the tangent to the manifold at the origin. We denote $w$ as the unit twist corresponding to the displacement $\hat{q}$, that is,

$$w = \frac{\Delta\theta}{2}(s + \epsilon(s^0 + hs)),$$

\hspace{1cm} (1)
for a displacement

\[
\hat{q} = \cos \frac{\Delta \theta}{2} - \epsilon \frac{\Delta t}{2} \sin \frac{\Delta \theta}{2} + (\sin \frac{\Delta \theta}{2} + \epsilon \frac{\Delta t}{2} \cos \frac{\Delta \theta}{2})S,
\]

with \( S = s + \epsilon s^0 \) being the Plucker coordinates of the screw axis of the displacement and \( h = \frac{\Delta t}{\Delta \theta} \) being the pitch, calculated as the ratio of the slide along the axis, \( \Delta t \) and the rotation about the axis \( \Delta \theta \).

The finite motion can be obtained as the exponential of the finite twist,

\[
\hat{q} = e^{\omega},
\]

Similarly, the logarithmic function performs the mapping of the displacement to the finite twist. The strategy used here is to perform the computations in the Lie algebra space and use the exponential mapping to find the corresponding displacements.

### 3.3 Motion Data Analysis

Using the logarithm and exponential mappings, techniques that require the calculation of distances can be applied to motion data. Principal Component Analysis (PCA) is able to eliminate the small, parasite motion always present in human data. Figure 7 shows the original trajectory of a human arm and the data recovered when considering only three principal components for the data.

Similarly, a desired number of positions can be extracted from the data that represent the clusters in the data. A few different techniques, such as k-means or EM algorithm, can be used here. Figure 7 shows the result of extracting three positions from a finger trajectory.

Extracting a small set of positions to perform exact kinematic synthesis is less computationally costly than performing a big optimization over the whole motion trajectory.
4 STRUCTURAL SYNTHESIS

Given the large number of topologies available for multi-fingered hands, it is necessary to derive some criterion to select the most suited topologies for a given task. While matching a workspace shape to a topology is still far from being accomplished except for simple cases [2], some bounding on the candidate topologies can be based on the number of positions and the number of fingertips. The selection criteria is solvability, that is, the ability of the topology for being synthesized, reaching the workspace defined by the positions and without overconstraining any of its parts.

Algorithms have been developed [17] to find solvable topologies for a defined task. The task is assumed to be in the $SE(3)$ group of rigid motion, and the goal is to find all topologies that can be paired with the task for dimensional synthesis, given a set of user-defined restrictions. Those extra conditions bound a problem that otherwise may not be bounded, depending on the topology [9].

User-defined inputs are the number of positions of the task $m$, the number of end-effectors $b$ and the total number of edges of the graph $e$. The output is the set of topologies that meet the solvability criteria subject to the user-defined conditions. The formula in Eq.(6) is applied to do the enumeration,

$$\sum_{i=1}^{e} j_i = \frac{(m - 1) \times 6 \times b}{(m + 3)}$$  \hspace{1cm} (4)

where $\sum_{i=1}^{e} j_i$ represent the total number of joints and $j_i$ are the joints for the $i$-th edge or serial chain. The number of joints per edge has to be between 1 and 5 for synthesis purposes, as a serial chain of length 6 or higher does not impose any restriction on the motion.

The presented algorithm includes three steps.

- Find all possible joint arrays with length equal to $e$ and satisfying Eq.(4), with entries between 1 and 5. The order of this part is $O(10^{e-1})$.
- Find all possible tree structures (parent-pointer arrays $p$) which meet the input criteria, including the number of branches and number of edges.
- Combine and check: For all possible parent pointer arrays, check all combinations of joint array and parent-pointer array for solvability. If the topology is solvable, then it becomes a candidate topology.

As an example, Table 4 shows the number of topologies obtained when the input is the number of positions, number of end-effectors and total number of edges. Due to the high number of solvable candidate topologies obtained, these cannot be included in the examples.
5 DIMENSIONAL SYNTHESIS

Dimensional kinematic synthesis seeks to find the position of the joint axes for a given topology, in order for each of the end-effectors to perform the set of displacements of the task. In this section we present a summary of the design methodology; for details, see [13].

Consider a hand topology with $b$ branches and a total of $n_e$ joint axes $S_i$, and create the sets of ordered indices $B_j$ of joints belonging to the serial chain starting at the root and ending on end-effector $j$, for $j = 1, \ldots, b$. Given a simultaneous task for each fingertip, characterized by a set of $m_p$ finite positions $\hat{P}^{b}_{ik}$ and $m_v$ velocities $\dot{P}^{b}_{ik}$, kinematic synthesis is applied by equating the forward kinematics equations of each branch to the relative displacement of the corresponding fingertip. Similarly, velocities can be defined for some of those task positions,

$$\mathbf{F}(\mathbf{S}, \Delta \theta, \dot{\theta}) = \begin{cases} \hat{P}^{j}_{1k} - \prod_{i \in B_j} e^{\alpha_i^k} S^k_i, & k = 2, \ldots, m_p \\ \dot{P}^{j}_{k} - \sum_{i \in B_j} \dot{\theta}_{i}^k S^k_i, & k = 1, \ldots, m_v \end{cases}$$

(5)

where $S^k_i$ is the $i^{th}$ joint axes when moved to position $k$. This yields a total of $6(m - 1)b$ independent equations to be simultaneously solved. In most of the cases, this set of equations is solved using numerical methods to obtain a kinematic design.

5.1 Solvability of Tree Topologies for Exact Dimensional Synthesis

We define a kinematic chain as solvable if we can find a positive rational number of positions for which the exact dimensional synthesis yields a finite number of solutions.

In the case of serial chains the solvability problem is trivial. When dealing with tree topologies, the task sizing must be done so that the system of equations can be solved simultaneously while not overconstraining any of the branches.

Let $\mathbf{D}^e_j$ be an $e \times 1$ vector containing the joint degrees-of-freedom for each edge of the contracted graph, and $\mathbf{D}^s_j$ be the $e \times 1$ vector containing the number of structural parameters (four per joint in the general case) for each edge of the contracted graph. Denote as $\mathbf{D}^n_{ee}$ the $b \times 1$ vector containing the degrees-of-freedom of the space of each end-effector, and $\mathbf{D}^c_{ee}$ the $b \times 1$ vector with the number of additionally imposed constraints.

<table>
<thead>
<tr>
<th>Example</th>
<th>$m$</th>
<th>$b$</th>
<th>$e$</th>
<th>Solvable topologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ex. 1</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>219</td>
</tr>
<tr>
<td>Ex. 2</td>
<td>21</td>
<td>5</td>
<td>8</td>
<td>34548</td>
</tr>
</tbody>
</table>

Table 1: Examples of type synthesis
(if any) for each branch. Define the vectors $B_i$ as a $b_i \times 1$ vector of ones corresponding to the branches for subgraph $i$, and $E_i$ as an $e_i \times 1$ vector of ones for the edges in the subgraph considered. Those can be easily computed using the end-effector path matrix and the incidence matrix of the graph. There are $2^{b_i} - 1$ possible subgraphs for any given rooted tree graph, including the overall graph. The maximum number of positions for the subgraph is given by

$$m_i = \frac{D_s^c \cdot E_i - D_n^a \cdot B_i}{D_{ee}^n \cdot B_i - D_i^e} + 1.$$  

(6)

In addition to these, all different and non-isomorphic subgraphs that appear when exchanging the root node with each of the end-effectors need to be considered, see [9].

As a summary, an overall solution exists only when considering the solvability of all root-to-end-effector subgraphs, including root and end-effector switching. In this case, considering $m_i$ as the number of positions for exact synthesis for a subgraph $i$ with $i \in S$ the set of all possible different end-effector subgraphs up to isomorphism, and $m$ for the overall graph, the topology is solvable if

1. $m \in \mathbb{Q}^+$
2. $m \leq m_i, \forall m_i \in \mathbb{Q}^+, i \in S$

In the case of a subgraph containing $c$ branches and being solvable for $m_i = m$ positions, that subgraph can be solved separately, which eliminates exactly $6c(m - 1)$ equations and the same number of unknowns, so that the rest of the graph can be solved a posteriori.

5.2 Numerical Solution

Both the structural and the dimensional synthesis have been implemented in a solver. The structural synthesis follows a straightforward enumeration process and the dimensional synthesis is solved using a hybrid solver that returns a single solution for each run. This software is freely available at the project webpage [1].

6 EXAMPLE

Let us consider a hand task that can be defined with five positions. For this task we can use a minimum of three fingers and a maximum of five fingers, for grasping and manipulation purposes. Three fingers are sufficient for stable grasping but adding the extra two fingers may help in some manipulation strategies.

We start the design process defining $m = 5$ number of task positions, $b = 3$ to $b = 5$. Applying the searching and solvability algorithms, we find, for $b = 3$, a total of 57 solvable topologies. For $b = 4$ and $b = 5$ we find a total of zero solvable topologies. Table 2 shows the solutions of the type synthesis stage. It is not possible to include the complete list
Table 2: Solvable topologies for 5 positions, with 3 to 5 fingertips

<table>
<thead>
<tr>
<th>Fingers</th>
<th>Number of edges</th>
<th>Topologies</th>
<th>Parent-pointer</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b = 3$</td>
<td>$e = 3$</td>
<td>1</td>
<td>${0,0,0}$</td>
</tr>
<tr>
<td></td>
<td>$e = 4$</td>
<td>23</td>
<td>${0,0,1,1}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>${0,1,1,1}$</td>
</tr>
<tr>
<td></td>
<td>$e = 5$</td>
<td>33</td>
<td>${0,1,1,2,2}$</td>
</tr>
</tbody>
</table>

of topologies; the parent-pointer representation is included but those expand in many different topologies for the number synthesis (number of joints per edge).

The simplest solvable topology able to perform this task has parent-pointer array $p = \{0,0,0\}$ and joint array $j = \{3,3,3\}$, that is, a $0-(3,3,3)$ topology with three R-R-R fingers and no wrist. One of the most complex topologies is the $3-(2,1-(2,1))$ topology, with $p = \{0,1,1,2,2\}$ and $j = \{3,2,1,2,1\}$. Out of these 57 topologies, we select the topology with $p = \{0,1,1,2,2\}$ and $j = \{2,1,2,2,2\}$, corresponding to the $2-(1-(2,2),2)$ hand, with one R joint at the wrist spanning two fingers, the first one spanning two more fingers for a total of three end-effectors. Figure 8 shows the graph of this topology.

![Figure 8](image)

**Figure 8:** Graph of the selected topology

We perform dimensional synthesis for this topology with five random finite displacements. The resulting set of equations from Eq.(5) consists of 114 highly nonlinear equations in 90 unknowns, 72 of which are independent. The results of five runs with different initial conditions are presented in Table 3. All five obtained solution were feasible and different, which makes us infer that there will be a big number of solutions for this problem.

The output of the kinematic synthesis stage is to be used as the input for a detailed design, using computer-aided tools.
Table 3: Dimensional synthesis solving results

<table>
<thead>
<tr>
<th>Run</th>
<th>Final error</th>
<th>Iterations</th>
<th>Time to solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$1.15 \times 10^{-13}$</td>
<td>1</td>
<td>12 sec.</td>
</tr>
<tr>
<td>2</td>
<td>$6.0 \times 10^{-12}$</td>
<td>2</td>
<td>10 sec.</td>
</tr>
<tr>
<td>3</td>
<td>$1.4 \times 10^{-13}$</td>
<td>1</td>
<td>5 sec.</td>
</tr>
<tr>
<td>4</td>
<td>$2.0 \times 10^{-13}$</td>
<td>9</td>
<td>29 sec.</td>
</tr>
<tr>
<td>5</td>
<td>$1.0 \times 10^{-13}$</td>
<td>1</td>
<td>7 sec.</td>
</tr>
</tbody>
</table>

7 CONCLUSIONS

In this article we present an overview of the kinematic synthesis theory applied to the design of robotic hands. This includes task definition, enumeration of topologies, structural or type synthesis, and dimensional synthesis. The synthesis results have been implemented in a solver that is freely available. The objective is to provide with a tool able to help in the preliminary design of robotic hand for specific grasping and manipulation tasks. Subsequent steps of optimization and detailed design are necessary to obtain a final design.

REFERENCES


BLADE CUTTING SIMULATION WITH CRACK PROPAGATION THROUGH THIN-WALLED STRUCTURES VIA SOLID-SHELL FINITE ELEMENTS IN EXPLICIT DYNAMICS

ALDO GHISI, FEDERICA CONFALONIERI AND UMBERTO PEREGO

Dipartimento di Ingegneria Civile e Ambientale, Politecnico di Milano
Piazza Leonardo da Vinci 32, 20133, Milano (ITALY)
aldo.ghisi@polimi.it, federica.confalonieri@polimi.it, umberto.perego@polimi.it

Key words: Solid Shell Finite Elements, Selective Mass Scaling, Layered Thin-Walled Structures, Explicit Dynamics, Blade Cutting.

Summary. To simulate the crack propagation due to blade cutting of a thin-walled shell structure, we propose a numerical technique based on solid shell finite elements and explicit time integration. The limitation on the critical time step due to the small thickness along the out-of-plane direction is overcome through a selective mass scaling, capable to optimally define the artificial mass coefficient for distorted elements in finite strains: since the selective scaling cuts the undesired, spurious contributions from the highest eigenfrequencies, but saves the lowest frequencies associated to the structural response, and since the method preserves the lumped form of the mass matrix, the calculations in the time domain are conveniently speeded up. The interaction of the cutting blade with the cohesive process zone in the crack tip region is accounted for by means of the so-called directional cohesive interface concept. Unlike in previous implementations, through-the-thickness crack propagation is also considered. This is of critical importance in particular in the case of layered shells, where one solid-shell element per layer is used for the discretization in the thickness direction and it is a necessary ingredient for future possible consideration of delamination processes. We show by applying the proposed procedure to the cutting of a thin-walled laminate used for packaging applications that this is a promising tool for the prediction of the structural response of thin-walled structures in the presence of crack propagation induced by blade cutting.

1 INTRODUCTION

In this paper we consider the simulation of the opening process of a thin laminate membrane sealing a carton package by means of cutting blades. This problem is actually challenging from a computational point of view, since the simulation must account for fracture initiation and propagation through a thin-walled laminate. The nonlinear material behaviour of the single layer and the small laminate thickness make recently introduced [1-2] approaches, based on solid-shell elements particularly attractive. Moreover, the dynamic nature of the fracture process and the severe nonlinearity due in particular to the interaction
with the cutting blade suggest to use an explicit dynamics time integration scheme. These issues require a detailed analysis of numerous numerical problems, whose solution is briefly recalled in the following sections. The simulation approach considered here is based on the directional cohesive element concept, first proposed in [14] to simulate blade-induced crack propagation in classical elastic shell elements and later reformulated for application to solid-shell elements [3]. Directional cohesive elements are special interface elements, to be interposed between adjacent shell elements, capable to account for the interaction between the cutting blade and the cohesive process zone. Section 2 briefly recalls the adopted solid-shell finite element formulation, section 3 details the selective mass scaling approach; in section 4 we sketch the ideas, i.e. the directional cohesive elements, at the base of the interaction between the cutting blade and the fracture resistance of the laminate, and in section 5 we show the application of the modelling to an opening laminate in the packaging industry.

2 SOLID SHELL ELEMENT FORMULATION

Solid-shell finite elements, unlike standard, structural shell elements make use of only translational degrees of freedom. Because of the three-dimensional kinematic formulation, the constitutive laws used for continuum, three-dimensional finite elements can be adopted as well, including direct consideration of thickness deformation. Unlike standard finite elements, solid-shell elements can be conveniently used for the discretization of layered thin-walled structures, by simple stacking a number of solid-shell elements, typically one or more per layer, one on top of the other through the shell thickness. In the following we choose the Q1STS finite element proposed by Schwarze and Reese [1-2], an eight-node hexahedron including a reduced integration and hourglass control, recently adapted to explicit dynamics in [6]. Solid-shell elements are affected by numerical problems, such as Poisson, volumetric, curvature and shear locking and by hourglassing, the latter due to reduced integration which is an important ingredient for an efficient explicit time integration scheme. In Schwarze and Reese implementation volumetric and Poisson’s locking are dealt with by the enhanced assumed strain (EAS) approach, while the assumed natural strain (ANS) concept is used for the transverse shear and curvature thickness locking. Hourglassing is controlled by adding suitably constructed artificial stiffness.

3 SELECTIVE MASS SCALING

In explicit structural dynamics the advantage of algorithm simplicity (there are no linear systems to be solved) is balanced by the small allowed stable time step; in the case of the widely diffused central difference scheme, the time step $\Delta t$ is dependent on the maximum eigenfrequency $\omega_{\text{max}}$ of the finite element mesh according to

$$\Delta t \leq \frac{2}{\omega_{\text{max}}}$$

(1)

in the undamped case [8]. $\omega_{\text{max}}$ is in turn conservatively bounded by the maximum eigenfrequency of the single finite element $\omega^e_{\text{max}}$:

$$\omega_{\text{max}} \leq \max_e(\omega^e_{\text{max}}).$$

(2)

Moreover, because of the large difference between the thickness and the in-plane dimensions
of a shell, the Courant-Friedrichs-Lewy condition, limiting the time step to a fraction of the element traversal time of an elastic wave, leads to very small time increments in explicit dynamics. To overcome this severe limitation, the selective mass scaling concept can be pursued: the idea is to artificially alter the solid-shell element mass matrix in order to scale down the highest eigenfrequencies with small or negligible changes to the lowest ones. There are several approaches to achieve this goal (see e.g. [13,15,16,17,18,19]), but most of these ideas make the scaled mass matrix non diagonal, an undesirable outcome for explicit time integration schemes. Here we follow the approach shown in the series of papers [4,6,9,10], where the relative motion between the lower and upper shell surfaces of a 8-node three-linear, solid-shell element is conveniently penalized so that the smallness of the thickness is not determining the time step anymore. The approach maintains the translational rigid body inertia, while the rigid body rotational inertia is modified; therefore, it should be used with care when a significant component of the motion is a pure rigid body rotation. Besides this drawback, we emphasize that the method works for regular and distorted finite elements in finite strains [4].

3.1 Principles of the proposed selective scaling approach
To summarize our selective mass scaling approach we use here the notation described in [4]: it starts by introducing the splitting of the element nodal coordinates according to the upper and lower surface (see figure 1):

$$ X = \begin{bmatrix} X_{1-4} \\ X_{5-8} \end{bmatrix}. \tag{3}$$

![Figure 1: Local numbering adopted for the solid-shell element.](image)

Then, the coordinates of the corner nodes at the middle surface and the corner fibers, i.e. the segments connecting the corresponding nodes of the upper and lower surface, are defined:

$$ X^m = \frac{X_{5-8} + X_{1-4}}{2} \quad \Delta X = \frac{X_{5-8} - X_{1-4}}{2} \tag{4a,4b} $$

and these items are collected in the transformed coordinate vector

$$ \tilde{X} = \begin{bmatrix} X^m \\ \Delta X \end{bmatrix} \tag{5}$$

which is related to the original coordinates via the linear transformation
\[ X = Q \ddot{x} \quad Q = \begin{bmatrix} \mathcal{I} & -\mathcal{I} \\ \mathcal{I} & \mathcal{I} \end{bmatrix} \]

where \( \mathcal{I} \) is the 12x12 identity matrix. A similar transformation can be followed also for
displacements, velocities, and, finally, accelerations, which in particular read:
\[
a = \begin{bmatrix} a_{1-4} \\ a_{5-8} \end{bmatrix}, \quad \hat{a} = \begin{bmatrix} a^m \\ \Delta a \end{bmatrix} \quad a = Q\hat{a}.
\]

The principle of virtual work for the motion of an undamped system can be therefore
transformed as follows:
\[
\delta \hat{a}^T_e \overline{M}_{e, \text{lumped}} \hat{a}_e = \delta \hat{a}^T_e \hat{f}_e
\]
with \( \delta \hat{a}_e \) equal to the transformed virtual acceleration of the nodes of element \( e \), \( \hat{f}_e = \hat{f}_e^{\text{ext}} - \hat{f}_e^{\text{int}} \) the difference between external and internal (transformed) equivalent nodal forces for
the single FE, and \( \overline{M}_{e, \text{lumped}} \) the lumped form of the selectively scaled mass matrix, namely
\[
\overline{M}_{e, \text{lumped}} = \begin{bmatrix} m_{u} + m_{l} & 0 \\ 0 & \alpha^e (m_{u} + m_{l}) \end{bmatrix}
\]

\( m_{u} \) and \( m_{l} \) being 12x12 diagonal matrices collecting the element nodal masses at the
upper and lower surface nodes, respectively, and \( \alpha^e \) being the element scaling coefficient.
Notice that \( \alpha^e \) is applied only to the difference in nodal acceleration between the upper and
lower surface and not to the middle surface nodal accelerations that are responsible for the
element translational rigid body motion. In case of a single layer thin-walled structure, this
mass matrix evidently does not change the diagonal form of the problem and can be directly
inserted in the explicit solver routines written in the transformed variables. In case, instead, of
a layered structure, it is necessary to reverse to original coordinates; in fact, after assembly,
the nodal masses of adjacent layers belonging to the same multi-layer fiber sum up, thus
requiring to solve a small linear system. While a detailed presentation is present in [9], the
basic idea can be grasped by considering the three layers (labelled as \( a, b, c \)) example in
figure 2. For each fiber \( f \) in this case the problem to be solved is
\[
M_f^2 \alpha_f = f_f
\]

```
M_f^2 = \begin{bmatrix} m_{a}^{LL} & m_{a}^{LU} & 0 & 0 \\ m_{b}^{UL} & m_{a}^{LU} + m_{b}^{LL} & m_{b}^{LU} & 0 \\ 0 & m_{b}^{UL} & m_{b}^{LU} + m_{c}^{LL} & m_{c}^{LU} \\ 0 & 0 & m_{c}^{UL} & m_{c}^{LU} \end{bmatrix}
```

and \( m_{i}^{LL} = m_{i}^{LU} = \sum(1 + \alpha^e) (m_{i}^{up} + m_{i}^{low}) \), \( m_{i}^{LU} = m_{i}^{UL} = \sum(1 - \alpha^e) (m_{i}^{up} + m_{i}^{low}) \),
where, for each node, the sum operator has to be interpreted as the assembly over the layer
\( l=a,b,c \) of the elements sharing that node.
3.2 Optimal mass scaling factor

Unnecessary high values of the scaling coefficient $\alpha$ can overly modify the structural dynamical properties (e.g. rigid body rotational inertia) and it can be shown [9] that the dependence of the maximum eigenfrequency on the mass scaling coefficient actually shows a plateau for increasing $\alpha$. Therefore, it is necessary to determine the optimal scaling: this problem has been solved in [10] for parallelepiped elements and in [4] for distorted elements, essentially by exploiting the equivalence between selective scaling and element thickness geometrical scaling. A good approximation of the optimal estimate for the mass scaling coefficient turns out to be:

$$\alpha_{\text{opt}} \approx \frac{L_{\text{min}}^2}{h_0^2}$$  \hspace{1cm} (12)

where $L_{\text{min}} = \min\{L_1, L_2\}$ is the minimum in-plane length and $h_0$ is the thickness of the solid-shell element. The estimate is actually exact for parallelepiped elements while only approximate for distorted elements (in [4] the exact solution is provided for interested readers). Moreover, this estimate is different for each layer of the laminate and for each element; by using the more restricted estimate, the time step depends only on the element in-plane dimensions and not on its thickness. To pass from the mass scaling coefficient stated in equation 12 to the time step estimate, we follow equations 1 and 2 where $\omega_{\text{max}}^2$ is conveniently bounded (details are given in [4]) by the solution of the element eigenvalue problem $f(\omega^2; \alpha_{\text{opt}}) = 0$ via a Newton-Raphson procedure.

4 DIRECTIONAL COHESIVE ELEMENTS

A sharp blade cutting a ductile (or not purely brittle) material can interact with the cohesive process zone during crack propagation. In this case, a classical implementation of the cohesive approach would lead to incorrect results unless an extremely fine, computationally prohibitive mesh, capable to resolve the blade radius of curvature, were used. The “directional” cohesive approach [3,14] is intended to provide a macroscopic but energetically consistent description of this interaction, allowing a crack to propagate along element interfaces where special string elements are inserted after node duplication. Under the assumptions of i) neglecting the bending strength of individual layers because of their small thickness, and ii) the independence of the specific surface work due to cutting from the crack
propagation mode (this implies that only fracture mode I is relevant, once the fracture criterion is met at a given node), one massless cable per layer is introduced between the separating element faces in correspondence of the element middle surfaces. The cable element is a geometric entity, capable to detect contact and to interact with the advancing cutting blade (figure 3a, 3b): the latter, in fact, modifies the original straight configuration of the cable, by generating a joint in correspondence of the contact point at the cutting edge (the possibility of contact with the lateral surfaces of the blade is excluded because of the assumed sharpness of the cutting blade), and allowing for a transmission of the cohesive force in the correct direction.

![Figure 3](image)

**Figure 3:** (a) blade cutting through the mesh, (b) interaction of the blade with the cable elements, and (c) “directional” cohesive forces.

Unlike in the classical cohesive approach, these forces have different directions on the two separating crack flanks, determined by the directions of the truss elements connecting the solid-shell nodes with the cable joint (see figures 3c and 4b). By assuming that the blade position is given at each time, then the contact point \( \mathbf{x}^m \) between blade and cable and the displacement discontinuities \( \delta^m \) are known at the end of each explicit time integration step, so that the cohesive forces become:

\[
\begin{align*}
  f_a^+ &= \frac{\|f_a\|}{\|\mathbf{x}^m - \mathbf{x}^{\text{mid}+}\|} (\mathbf{x}^m - \mathbf{x}^{\text{mid}+}) \\
  f_a^- &= \frac{\|f_a\|}{\|\mathbf{x}^m - \mathbf{x}^{\text{mid}-}\|} (\mathbf{x}^m - \mathbf{x}^{\text{mid}-})
\end{align*}
\]  

(13a,13b)

where the subscripts “+” and “-” denote opposite crack flanks, \( \mathbf{x}^m \) is the contact point between the cable and the blade and \( \mathbf{x}^{\text{mid}} \) is the point of the crack flank where the cable is attached in correspondence of the element middle surface.

The forces \( f_a \) are calculated from the cohesive model described in [20], assuming a linear softening law relating the opening traction \( T \) to a scalar measure of the crack opening \( \ell \) and a cohesive potential \( G(\ell, q) \) with \( q \) an internal variable, shown in figure 4a for the simple case of a linear softening law, i.e. the maximum attained crack opening, also interpreted as a progressive damage inside the process zone.
A softening function \( g(q) \) is also introduced so that, in the linear softening case, the cohesive law is written as follows:

\[
G = G_f + \frac{1}{2} T_0 \ell \cr
G_f = \frac{1}{2} T_0 \ell \cr
T = \frac{\partial G}{\partial \ell}(\ell, q) = \begin{cases} \frac{T_0}{q} \left(1 - \frac{q}{\ell} \right) \ell = \frac{g(q)}{q} & \text{when } \ell \leq q, \dot{q} = 0 \\
\frac{T_0}{q} \left(1 - \frac{q}{\ell} \right) \ell = \frac{g(q)}{q} & \text{when } q, \dot{q} \geq 0 \end{cases}
\]

It is therefore obtained \( f_\alpha = T_\alpha^{\text{mid}} A_\alpha \), while \( \ell = \|x^m - x^{\text{mid}+}\| + \|x^m - x^{\text{mid}^-}\| \) is the length of the cable. The parameter \( T_0 \) is set equal to the component of the stress tensor in the direction normal to the opening faces of the element in the continuum model where the fracture criterion is met; \( \ell \cr \), is, instead, assumed to be a material parameter.

An equivalent plastic strain criterion, with a Heaviside function assuring that only tensile stresses contribute to the fracture, is adopted:

\[
\varepsilon^p_{\ell \cr} = \int_{0}^{\varepsilon^p_{\ell \cr}} H(\sigma_H) \, d\varepsilon^p
\]

where \( \varepsilon^p_{\ell \cr} \) is the plastic strain value at fracture, \( \sigma_H \) is the hydrostatic stress accounting for the influence of triaxiality, and the crack direction coincides with the element edge closer to the maximum principal stress at the node where the criterion has been met.

5 RESULTS

The proposed approach has already proven itself promising in several single-layer applications [3,6]. Here we show an application to a multi-layered structure. The extension is not trivial because of the much higher complexity of the kinematics of the fracture opening during the through-the-thickness crack propagation.

A thin laminate membrane closes the carton package for fluid aliments [7] shown in figure 5. The circular hole to be produced through the applied cap equipped with high density polyethylene (HDPE) cutting teeth has the diameter of 15 mm: it is obtained with a torque applied by the user to the screw thread. In the remainder of the package the laminate constituting the hole does not contain paper, but only aluminum and low density polyethylene.
Aldo Ghisi, Federica Confalonieri and Umberto Perego

(LDPE); this laminate has a thickness ranging from 70 to 78 μm, with the aluminum core 6-9 μm thick, while a LDPE layer coat this interior core from both sides.

![Figure 5](image1)

**Figure 5:** (a) package with hole for opening, (b) package with the screw thread applied, (c) detail of the cutting teeth (upside down view).

In the modelling we assume a circular membrane with radius equal to 10.2 mm and thickness 0.074 mm. The 14% larger diameter accounts for the slackness of the actual membrane due to the manufacturing process. As shown in figure 5c, the cutting tool (referred to as the “blade” from now on) has a radius of 9.4 mm and its edges present a curvature radius of 0.1 mm (see also figure 6b). We consider the blade (figure 6a) as a rigid body whose motion is given: it has a rotation around its vertical axis and a translation along the same axis penetrating the three-layer membrane shown in figure 6b.

![Figure 6](image2)

**Figure 6:** (a) FE model of blade and membrane; (b) detail of the three layers membrane.

The composite laminate in this example is assumed as an equivalent and homogenized membrane with Young modulus $E=176$ GPa, Poisson’s ratio $v=0.3$ and density $\rho=2 \times 10^{-9}$ Ns²/mm⁴. A hardening behavior is adopted according to the yielding function

$$
\sigma_{y}(\varepsilon) = \sigma_{y0} + Q \left(1 - \exp(-\zeta \varepsilon)\right)
$$

(27)

where $\sigma_{y0}=11$ MPa is the initial stress, the saturation parameter $Q$ is equal to 40 MPa, and the hardening exponent $\zeta=6.63$. The chosen fracture parameters are $G_{f}=30$ N/mm and $T_{0}=18$ MPa. The blades produces several cuts along their circular path as shown in figure 6a; the cut appearance (figure 7a) is in good qualitative agreement with experimental observations of the opening process, as much as the torque vs cap rotation responses shown in figure 7b. This comparison shows that, neglecting the initial mismatch between the two curves, due to the
lack of friction imposed on the rotating teeth before contact with the laminate, the agreement is reasonable.

Figure 7: (a) upside down view of the membrane cut by the blades; (b) comparison between numerical and experimental torque during the rotation.

6 CONCLUSIONS

A simulation approach for the numerical evaluation of the opening process of a thin-walled laminate has been presented. It makes use of solid-shell elements enriched to correct known locking issues, in a finite strain formulation. To conveniently exploit the advantages of an explicit time integration scheme in structural dynamics, an effective selective mass scaling technique has been implemented, whereby only the highest eigenfrequencies responsible of the smallness of the time step are reduced. To correctly describe the interaction between the blade and the fracturing multi-layered thin-walled membrane, interface directional cohesive elements have been used. We have shown that several layers can be used for the description of the laminate, and fracture initiation and propagation in multi-layered structures is well represented; in the future the possible delamination between layers will be considered.

ACKNOWLEDGMENTS

We are grateful for the support given by Tetra Pak Packaging Solutions.

REFERENCES


STABILITY AND ERROR ESTIMATE OF A COHESIVE ZONE MODEL IMPLEMENTED USING THE AUGMENTED LAGRANGIAN METHOD.

Nicolás Agustín Labanda¹, Sebastián Miguel Giusti², Bibiana María Luccioni¹

¹Universidad Nacional de Tucumán and CONICET
Av. Independencia 1800, San Miguel de Tucumán, Argentina
nlabanda@facet.unt.edu.ar

²Universidad Tecnológica Nacional (FRC) and CONICET
Maestro M. López esq. Cruz Roja Argentina, Córdoba, Argentina
sgiusti@civil.frc.utn.edu.ar

Key words: Cohesive Zone Model, Fracture, Augmented lagrangian method.

Abstract. Since the origins of the cohesive zone model (CZM) proposed by Dugdale [1] and later by Barenblatt [2], the approach has been increasingly used and studied in computational mechanics community and several traction-separation criteria have been proposed to analyse damage in different kinds of materials. The classical way to implement those CZM is straightforward considering a fracture equilibrium term in the global solid equilibrium equation as an intrinsic behaviour. In recent years new approaches like enrichment-embedded kinematics, discontinuous Galerkin methods, isogeometric analysis, lagrange multipliers based formulations have been successfully used for quasi-static and dynamic fracture simulation. A CZM implemented in an augmented lagrangian formulation based on the model proposed by Lorentz [3] is developed and analyzed in this paper. This method is able to deal with unilateral contact and cohesive forces via a supplementary variable that enforces the jump displacement in so-called collocation points. It represents a suitable tool to study the debonding phenomena in composites with strongly different stiffness, avoiding ill-conditioning problems associated with penalty methods. The model stability and an error estimation following Brezzi theorem [4] are discussed. Some numerical examples that show the ability of this approach to capture inclusions debonding are included in the paper.

1 VARIATIONAL FRAMEWORK

Let \( \mathcal{L}_u : \mathcal{U} \to \mathcal{R} \) be a system potential expressed in equation 1 as a function of body displacements \( u \) corresponding to a elastic volume \( \Omega \) with a fracture \( \Gamma \), Neumann
boundary $\Gamma^N$ and Dirichlet boundary $\Gamma^D$ presented in the figure 1.

![Figure 1: Domain decomposition of $\Omega$ considering a cohesive fracture $\Gamma$.](image)

\[
\mathcal{L}_u(u) = \mathcal{L}^B(u) + \mathcal{L}^F([u])
\]  

(1)

where

\[
\mathcal{L}^B(u) = \frac{1}{2} \int_{\Omega \setminus \Gamma} \sigma(x,t) : \nabla u(x,t) \ d\Omega - \int_{\Omega \setminus \Gamma} b(x,t) \cdot u(x,t) \ d\Omega - \eta \int_{\Gamma^N} \overline{p}(x,t) \cdot u(x,t) \ d\Gamma^N
\]

(2)

\[
\mathcal{L}^F([u]) = \int_{\Gamma} \Psi([u],\kappa) \ d\Gamma
\]

being the set of kinematically admissible displacement $\mathcal{U}(\Omega \setminus \Gamma)$ given by:

\[
\mathcal{U}(\Omega \setminus \Gamma) = \{ u \in H^1(\Omega \setminus \Gamma) \land [u] \cdot n \geq 0 \in H^2(\Omega \setminus \Gamma) : u|_{\Gamma^D} = \overline{u} \}
\]

(3)

Furthermore, $[\bullet] = (\bullet)|_{\Gamma^+} - (\bullet)|_{\Gamma^-}$ represents the jump of the field $(\bullet)$ over the domain $\Gamma$, $\Psi$ is the cohesive energy density on the fracture $\Gamma$, $\eta \in \mathbb{R}$ is a load parameter which acts proportionally to a unit load $\overline{p} \in L^2(\Gamma^N)$. $\kappa$ is an internal variable on the interface, $b \in L^2(\Omega \setminus \Gamma)$ is the volumetric force, and finally $\sigma$ is the Cauchy tensor.

Equilibrium problem can be stated as follows:

\[
\min_{u \in \mathcal{U}} \mathcal{L}_u(u)
\]

(4)

Following Lorentz [3], an indicator function is included in the pseudo-potential over the crack $\Psi$ to avoid penetration between crack lips.
\[ \Psi ([u], \kappa) = I_{\mathbb{R}^+} ([u]_n) + \psi ([u], \kappa) \quad \text{with} \quad \psi : \mathbb{R}^+ \rightarrow \mathbb{R} \quad (5) \]

Where the operator \( I_{\mathbb{R}^+} ([u]_n) \) is detailed in equation 6 in terms of the crack normal vector \( n, \psi \) is the cohesive potential.

\[ I_{\mathbb{R}^+} ([u]_n) = \begin{cases} 0 & \text{if} \quad [u] \cdot n \geq 0 \\ +\infty & \text{if} \quad [u] \cdot n < 0 \end{cases} \quad (6) \]

The functional \( L_u \) is reformulated using an augmented Lagrangian method and later, the decomposition coordination method (or collocation points method) proposed by Fortin [9]. In this way, a decoupling of the formulation into a global linear problem, and a family of local problems solved in the collocation points is carry out. This two set of problems are coordinated by Lagrange multipliers giving a classical saddle point problem. To this end, \( \delta \) a supplementary variable subjected to the restriction \([u] - \delta = 0\) is introduced. Then, equation 1 is rewritten as:

\[
\begin{aligned}
\min_{(u, \delta) \in \mathcal{V} \times X} & \quad L^B (u) + L^F (\delta) \\
\text{subject to} & \quad [u] - \delta = 0
\end{aligned}
\quad (7)
\]

The augmented Lagrangian \( L : \mathcal{V} \times \mathbb{R} \times \mathcal{W} \rightarrow \mathbb{R} \) associated to problem 7 is:

\[
L (u; \delta; \lambda) = L^B (u; \delta) + L^F (\delta) + L^C (u; \delta; \lambda)
\quad (8)
\]

where:

\[
L^B (u, \delta) = L^B (u) + \frac{\gamma}{2} ([u] - \delta)^2 \, d\Gamma
\]

\[
L^F (\delta) = \int_{\Gamma} \Psi (\delta, \kappa) \, d\Gamma, \quad L^C (u; \delta, \lambda) = \int_{\Gamma} \lambda \cdot ([u] - \delta) \, d\Gamma
\quad (9)
\]

Note that the second term in equation 8 corresponding to the cohesive energy density \( \Psi \) becomes a function of the supplementary variable, which is controlled in the third term by the Lagrange multipliers \( \lambda \) fixing in a weak sense the restriction \([u] - \delta = 0\). The stationary point of the augmented Lagrangian is:

\[
\min_{(u, \lambda, \delta) \in \mathcal{V} \times \mathbb{R} \times X} L (u; \delta; \lambda) = \partial_u L \cdot \delta u + \partial_\lambda L \cdot \delta \lambda + \partial_\delta L \cdot \delta \delta = 0
\quad (10)
\]

Each term represents the equilibrium equations of the fracture problem. The operator \( \partial_\delta (\cdot) \) represents the sub gradient or generalized derivative of the functional \( (\cdot) \) respect to
Only when \( (\cdot) \) is differentiable, the sub gradient contains only one element equal to the partial derivative \( \frac{\partial (\cdot)}{\partial \delta} \) (see work [8]).

\[
\mathcal{G}_1 (\mathbf{u}; \delta; \lambda) = \int_{\Omega \setminus \Gamma} \mathbf{\sigma} (x, t) : \nabla \delta \mathbf{u} (x, t) \, d\Omega + \int_{\Gamma} [\lambda + \gamma \cdot ([\mathbf{u}] - \delta)] [\delta \mathbf{u}] \, d\Gamma = \\
= \eta \int_{\Gamma_N} \mathbf{p} (x, t) \cdot \delta \mathbf{u} \, d\Gamma_N + \int_{\Omega \setminus \Gamma} \mathbf{b} (x, t) \cdot \delta \mathbf{u} \, d\Omega
\]

\[
\mathcal{G}_2 (\mathbf{u}; \delta; \lambda) = \int_{\Gamma} ([\mathbf{u}] - \delta) \cdot \delta \lambda \, d\Gamma = 0
\]

\[
\mathcal{G}_3^* (\mathbf{u}; \delta; \lambda) = \int_{\Gamma} [\partial \delta \Psi (\delta, \kappa) - \lambda - \gamma ([\mathbf{u}] - \delta)] \cdot \delta \delta \, d\Gamma = \\
= \int_{\Gamma} [\mathbf{t} - \lambda - \gamma ([\mathbf{u}] - \delta)] \cdot \delta \delta \, d\Gamma = 0 \quad \mathbf{t} \in \partial \delta \Psi (\delta, \kappa)
\]

With \( \delta \mathbf{u} \in \mathcal{V} \), space of kinematically admissible displacement variations, \( \delta \lambda \in \mathcal{W} \) space of kinematically admissible Lagrange multipliers variations and in the same way, \( \delta \delta \in \mathcal{X} \) space of kinematically admissible supplementary variable variations:

\[
\begin{align*}
\mathcal{V} (\Omega \setminus \Gamma) &= \left\{ \mathbf{v} \in H^1 : ([\mathbf{v}] \in H^\frac{1}{2} : \mathbf{v} |_{\Gamma_D} = 0 \right\} \\
\mathcal{W} (\Omega \setminus \Gamma) &= \left\{ \mathbf{w} \mid \mathbf{w} \in H^{-\frac{1}{2}} \right\} \\
\mathcal{X} (\Omega \setminus \Gamma) &= \left\{ \mathbf{x} \mid \mathbf{x} \in H^\frac{1}{2}_0 : \mathbf{x} \cdot \mathbf{n} \geq 0, \forall \mathbf{x} \in \Gamma \right\}
\end{align*}
\]

\( \mathcal{G}_3^* \) is a non differentiable equation which relates the constitutive behaviour of the variable \( \mathbf{t} \in \partial \delta \Psi \) with the Lagrange multipliers, from which a generalization of Hertz–Signorini–Moreau condition for contact problems is obtained [7]. By the coordination decomposition method, this functional allows to express the supplementary variable as a function of the displacements and Lagrange multipliers \( \hat{\delta} = \hat{\delta} (\mathbf{u}; \lambda) \).

### 2 ANALYSIS OF STABILITY AND A-PRIORI ERROR ESTIMATION

This section analyse the uniqueness of the solution in the direct minimization problem 4, and using the Brezzi theorem [4] the existence and uniqueness of the solution for the mixed formulation.

#### 2.1 Existence, uniqueness and equivalence between direct and mixed formulation.

Let’s define the space of kinematically admissible displacement \( \mathcal{K} \):

\[
\mathcal{K} = \{ \mathbf{v} \in H^1 : (J\mathbf{v}) \cdot \mathbf{n} \geq 0 \quad \text{in} \quad \Gamma \quad \text{and} \quad \mathbf{v} = 0 \quad \text{in} \quad \Gamma_D \} \subseteq \mathcal{V} \subseteq \mathcal{W}
\]
The apostrophe represents the dual space and the operator $J \in \mathcal{L}(\mathcal{V}, \mathcal{X})$ defined by $J : v \in \mathcal{V} \mapsto [v] = Jv \in \mathcal{X}$, maps bulk displacements in gaps onto fracture.

**Proposition 2.1.** Let $\mathcal{L}^B$ and $\mathcal{L}^F$ (equation 2) be continuously differentiable functionals in $\mathcal{V} (\Omega \setminus \Gamma)$ and $\exists k > 0 : \forall u, v \in \mathcal{V}, \| \partial \mathcal{L}^F (u) - \partial \mathcal{L}^F (v) \|_{\mathcal{X}'} \leq k \| u - v \|_{\mathcal{X}}$, then the functional $1$ is differentiable such that a solution $u$ of the problem $4$ satisfies:

\[
\langle \partial \mathcal{L}_u, v - u \rangle_{\mathcal{X}', \mathcal{X}} \geq 0, \quad \forall v \in \mathcal{X} \tag{14}
\]

Furthermore, if $\mathcal{L}^B$ is $\mathcal{V}$-elliptic with constant $\theta_B$ and considering a continuity constant $c_J$ for $J$ operator, it can be shown that $\exists \alpha > 0 \in \mathbb{R}$ such that $\mathcal{L}_u$ become strongly monotone, so the problem $4$ has unique solution.

**Proof.** $\mathcal{L}_u$ is differentiable by hypothesis so equation $14$ holds. The proof of uniqueness of problem $4$ follows:

\[
\langle \partial \mathcal{L}_u (v) - \partial \mathcal{L}_u (u), v - u \rangle_{\mathcal{X}', \mathcal{X}} \geq \langle \partial \mathcal{L}_u (v) - \partial \mathcal{L}_u (u), v - u \rangle_{\mathcal{X}', \mathcal{X}} + \langle \partial \mathcal{L}_u (Jv) - \partial \mathcal{L}_u (Ju), Jv - Ju \rangle_{\mathcal{X}', \mathcal{X}} \\
\geq \theta^B \| v - u \|_{\mathcal{X}}^2 + \theta^B \| v - u \|_{\mathcal{X}}^2 + \theta^B \| Jv - Ju \|_{\mathcal{X}}^2 \geq \theta^B \| v - u \|_{\mathcal{X}}^2
\]

if $\alpha = \theta^B - k c_J c_{\mathcal{X}} > 0$, then the solution of problem $4$ is unique. \qed

Parameter $\alpha$ can be related with physical properties of the numerical model, being $\theta^B$ proportional to the norm of the constitutive tensor in the bulk phase $\Omega$, the Lipschitz continuity constant $k$ grows with the decrease of cohesive forces. The continuity constant $c_J c_{\mathcal{X}}$ tends to zero when the norm of $\Gamma$ tends to zero.

In the following propositions, an extension of the results reported in [6] is performed, adapting those to cohesive fracture problems.

**Proposition 2.2.** Let $u \in \mathcal{X}$ be the unique solution of problem $4$, the augmented Lagrangian $8$ has an unique saddle point $(u, \delta, \lambda) \in \mathcal{V} \times \mathcal{X} \times \mathcal{W}$ such that:

\[
\mathcal{L} (u, \delta, \mu) \leq \mathcal{L} (u, \delta, \lambda) \leq \mathcal{L} (v, p, \lambda), \quad \forall (v, p, \mu) \in \mathcal{V} \times \mathcal{X} \times \mathcal{W} \tag{16}
\]

Moreover, the following relations can be proved:

\[
\lambda = -\sigma \cdot n, \quad \delta \cdot n = (Ju) \cdot n \geq 0 \quad \text{and} \quad \lambda = \partial_{\delta} \mathcal{L}^F \tag{17}
\]

when the derivative is defined (active cohesion).
Proof. (⇐ Suppose that \((u, \delta, \lambda) \in \mathcal{V} \times \mathcal{X} \times \mathcal{W}\) is a saddle point of \(L\), then \(u \in \mathcal{X} \subseteq \mathcal{V}\) is a minimum of functional 1 and consequently equations 17 holds).

Taking the left inequality of 16:

\[
\mathcal{L}(u, \delta, \mu) \leq \mathcal{L}(u, \delta, \lambda), \quad \forall \mu \in \mathcal{W}
\]

\[
\int_{\Gamma} \mu \cdot (Ju - \delta) \, d\Gamma \leq \int_{\Gamma} \lambda \cdot (Ju - \delta) \, d\Gamma, \quad \forall \mu \in \mathcal{W}
\]

If \(\mu = 2\lambda \in \mathcal{W}\), then:

\[
\int_{\Gamma} \mu \cdot (Ju - \delta) \, d\Gamma \geq \int_{\Gamma} \lambda \cdot (Ju - \delta) \, d\Gamma
\]

which yields \(\int_{\Gamma} (\mu - \lambda) \cdot (Ju - \delta) \, d\Gamma = 0 \quad \forall \mu \in \mathcal{W}\), so \(Ju = \delta\) over the fracture domain \(\Gamma\). By taking the left inequality:

\[
\mathcal{L}(u, \delta, \lambda) \leq \mathcal{L}(v, p, \lambda), \quad \forall (v, p) \in \mathcal{V} \times \mathcal{X}
\]

\[
\mathcal{L}^B(u) + \mathcal{L}^F(\delta) + \int_{\Gamma} \lambda \cdot ([u] - \delta) \, d\Gamma + \int_{\Gamma} \frac{\gamma}{2} ([u] - \delta)^2 \, d\Gamma
\]

\[
\leq \mathcal{L}^B(v) + \mathcal{L}^F(p) + \int_{\Gamma} \lambda \cdot ([v] - p) \, d\Gamma + \int_{\Gamma} \frac{\gamma}{2} ([v] - p)^2 \, d\Gamma
\]

Due to the indicator function in \(\mathcal{L}^F(\delta)\), the restriction \(\delta \cdot n \geq 0\) should be added, otherwise the inequality does not hold. If \(v \in \mathcal{X}\), and taking \(p = Ju\) over \(\Gamma\):

\[
\mathcal{L}^B(u) + \mathcal{L}^F(Ju) \leq \mathcal{L}^B(v) + \mathcal{L}^F(Jv) + \int_{\Gamma} \lambda \cdot ([v] - p) \, d\Gamma + \int_{\Gamma} \frac{\gamma}{2} ([v] - p)^2 \, d\Gamma
\]

Finally

\[
\mathcal{L}_u(u) \leq \mathcal{L}_u(v), \quad \forall v \in \mathcal{X}
\]

Being \(u \in \mathcal{X}\) the minimum of \(L_u\) and solution of the problem 4.

Let’s prove the first postulate in equation 17 by starting with equation 18, taking \(p = \delta\) and a vector \(v = u + t(w - u) \in \mathcal{V}\) with \(t \in [0, 1]\). Observe that the cohesive potential have no influence in this matter.
\[
\mathcal{L}^B(u + t(w - u)) - \mathcal{L}^B(u) + \int_\Gamma \lambda \cdot (t([w] - [u])) \, d\Gamma + \int_\Gamma \frac{\gamma}{2} (t([w] - [u]))^2 \, d\Gamma \geq 0
\]
\[
\mathcal{L}^B(u + t(w - u)) - \mathcal{L}^B(u) + \int_\Gamma \lambda \cdot ([w] - [u]) \, d\Gamma + \int_\Gamma t \frac{\gamma}{2} ([w] - [u])^2 \, d\Gamma \geq 0
\]

Taking the limit with \( t \to 0^+ \) in both sides
\[
\int_{\Omega \setminus \Gamma} \sigma(x, t) \cdot \nabla^s (w - u) \, d\Omega + \int_{\Gamma} \lambda \cdot J (w - u) \, d\Gamma \geq \eta \int_{\Gamma^N} \varphi(x, t) \cdot (w - u) \, d\Gamma^N + \int_{\Omega \setminus \Gamma} b(x, t) \cdot (w - u) \, d\Omega
\]

redefining \( v = (w - u) \), it can be proved:
\[
\int_{\Omega \setminus \Gamma} \sigma(x, t) : \nabla^s v \, d\Omega + \int_{\Gamma} \lambda \cdot J v \, d\Gamma = \eta \int_{\Gamma^N} \varphi(x, t) \cdot v \, d\Gamma^N + \int_{\Omega \setminus \Gamma} b(x, t) \cdot v \, d\Omega
\]

Applying Green’s theorem in the first term:
\[
\int_{\Omega \setminus \Gamma} \sigma(x, t) : \nabla^s v \, d\Omega = - \int_{\Omega \setminus \Gamma} (\text{div} \sigma) \cdot v \, d\Omega + \int_{\Gamma^N} (\sigma \cdot n) \cdot v \, d\Gamma^N + \int_{\Gamma} (\sigma \cdot n) J v \, d\Gamma
\]

replacing in equation 19, the relation that we are looking for \( \lambda = -\sigma \cdot n \) is obtained.

To prove the last relation in 17, take a vector \( p = \delta + t(\vartheta - \delta) \) with \( t \in [0, 1] \) and \( u = v \) in 18 to get the relation:
\[
\int_{\Gamma} (\partial_\delta \mathcal{L}^F - \lambda) \cdot (\vartheta - \delta) \, d\Gamma \geq 0, \quad \forall \vartheta \in \mathcal{X}
\]

hence, equation in 17 holds (if the derivative is defined).

(\( \Rightarrow \))

Let’s prove the uniqueness of the minimum of the augmented Lagrangian \( \mathcal{L} \), i.e. \( \exists! (u, \delta, \lambda) \in \mathcal{V} \times \mathcal{X} \times \mathcal{W} \) that verifies relations 16, with \( u \) the unique solution of the direct problem 4.

Taking an element \( (u, \delta, \lambda) \in \mathcal{V} \times \mathcal{X} \times \mathcal{W} \) with \( u \in \mathcal{V}, [u] \in \mathcal{X} \) minimum of \( \mathcal{L}_u \). Defining \( \delta = J u \in \mathcal{X} \), the left inequality in 16 is accomplished \( \forall \mu \in \mathcal{W} \). With \( u \) minimum of the direct problem, it can be shown from 14 and 19 that the relation \( \int_\Gamma \lambda \cdot [v] \, d\Gamma = 0, \quad \forall v \in \mathcal{V} \) holds and \( \int_\Gamma \lambda \cdot p \, d\Gamma \leq 0, \quad \forall p \in \mathcal{X} \) is satisfied because \( \lambda \) and \( p \) have the same direction but opposite senses. Then:
\[ L_u(u) \leq L_u(v), \forall v \in V \]
\[ L_u(u) \leq L_u(v) + \int_{\Gamma} \lambda \cdot ([v] - p) \, d\Gamma, \quad \forall (v, p) \in V \times X \]

So the right inequality expressed like 18 also holds.

### 2.2 Discretization of the saddle point problem.

Let \( \{\Omega_h\}_{h>0} \) be a mesh family that converges exactly to \( \Omega \setminus \Gamma \), where the parameter \( h>0 \) is the maximum element size, and \( \{\Gamma_h\}_{h>0} \) the set of mesh boundary. The spaces \( V_h, W_h \) are the approximations to the continuum spaces.

Motivated by the decomposition coordination method, the approximation of the jump is non conforming, so \( X_h \not\subset X \) do not hold in general because of the use of a discontinuous finite element discretization. The functional \( G^* \) in equation 11 is solved with fixed values of \( \lambda \) and \( u \), resulting in \( n_\Gamma \times n_c \) independent equations, where \( n_\Gamma \) is the number of interface elements and \( n_c \) the number of collocation points or quadrature points such that \( \delta_h \in \text{ker} G^*_3(u_h; \lambda_h) \subset X_h \).

\[ X_h = P_k^d(\Gamma_h) = \left\{ \delta_h \in L^2(\Gamma_h) : \forall I \in \Gamma_h, \delta|_I \in P_k \land \delta \cdot n \geq 0 \right\} \not\subset X \]  \hspace{1cm} (21)

\( P_k \) is the polynomial space with order less or equal to \( k \), \( I \) is a discrete element in \( \Gamma_h \) and the sub index \( d \) represents a discontinuous interpolation. The discrete conforming spaces for displacements and Lagrange multipliers are stated as follows:

\[ V_h = P^k_c(\Omega_h) = \left\{ v_h \in C^0(\Omega_h) : \forall E \in \Omega_h, v_h|_E \in P_k \land v_h|_{\Gamma^h} = 0 \right\} \subset V \]  \hspace{1cm} (22)

\[ W_h = P^k_c(\Gamma_h) = \left\{ \lambda_h \in C^0(\Gamma_h) : \forall I \in \Gamma_h, \lambda_h|_I \in P_k \right\} \subset W \]  \hspace{1cm} (23)

So, assuming a quadrature, the discrete augmented Lagrangian functional \( L_h : V_h \times X_h \times W_h \rightarrow \mathbb{R} \) is introduced

\[ L_h(u_h; \delta_h; \lambda_h) = L^G_h(u_h; \delta_h) + L^F_h(\delta_h) + L^C_h(u_h; \delta_h; \lambda_h) \]  \hspace{1cm} (24)

**Proposition 2.3.** If exists \( \alpha \in \mathbb{R} \) positive and independent of \( h \) such that the inf-sup condition is accomplished:

\[ \sup_{\mu_h \in W_h \atop \mu_h \neq 0} \frac{\int_{\Gamma_h} \mu_h \cdot (J_h v_h) \, d\Gamma_h}{\|v_h\|_V} \geq \alpha h^\frac{1}{2} \|\mu_h\|_W, \quad \forall \mu_h \in W_h \]  \hspace{1cm} (25)

then \( L_h \) admits a unique saddle point \( (u_h, \delta_h, \lambda_h) \in V_h \times X_h \times W_h \), i.e.:
\[ L_h(u_h, \delta_h, \mu_h) \leq L_h(v_h, \delta_h, \lambda_h) \leq L_h(v_h, p_h, \lambda_h), \quad \forall (v_h, p_h, \mu_h) \in \mathcal{V}_h \times \mathcal{X}_h \times \mathcal{W}_h \quad (26) \]

Moreover, the following relations hold:

\[
\int_{\Omega_h \setminus \Gamma_h} \sigma_h : \nabla \delta u_h \, d\Omega_h + \int_{\Gamma_h} \lambda_h \cdot (J_h \delta u_h) \, d\Gamma_h = L_h(\delta u_h), \quad \forall \delta u_h \in \mathcal{V}_h \quad (27)
\]

\[
\int_{\Gamma_h} (J_h u_h - \delta_h) \cdot \delta \lambda_h \, d\Gamma_h = 0, \quad \forall \delta \lambda_h \in \mathcal{W}_h \quad (28)
\]

\[
J_h u_h - \delta_h = 0, \quad \delta h \cdot n_h \geq 0, \quad \partial_n \mathcal{L}_h^F - \lambda_h = 0, \quad \forall c \in \Gamma_h \quad (29)
\]

where \( c \in \Gamma_h \) are collocation points and \( L_h(\delta u_h) = \eta \int_{\Gamma_h} p_h \cdot \delta u_h \, d\Gamma_h + \int_{\Omega_h \setminus \Gamma_h} b_h \cdot \delta u_h \, d\Omega_h \) in equation 27.

**Proof.** The first stage in the proof, shows the equivalence between equation 26 and the set of expressions 27-29. Equation 27 and 28 are obtained straightforwardly due the conformity in \( \mathcal{V}_h \subset \mathcal{V} \) and \( \mathcal{W}_h \subset \mathcal{W} \) when the left and right inequalities in 26 are taken. Motivated by the coordination decomposition method, the relations in 29 are calculated over the collocations points. The right inequality in 26 can be written as:

\[
\mathcal{L}_h^B(u_h) + \mathcal{L}_h^F(\delta_h) \leq \mathcal{L}_h^B(v_h) + \mathcal{L}_h^F(p_h) + \int_{\Gamma_h} \lambda_h \cdot [J_h(v_h - u_h) - p_h + \delta h] \, d\Gamma_h + \\
\gamma \int_{\Gamma_h} \frac{1}{2} [J_h(v_h - u_h) - p_h + \delta h]^2 \, d\Gamma_h, \quad \forall (v_h, p_h) \in \mathcal{V}_h \times \mathcal{X}_h \quad (30)
\]

the constraint \( \delta h \cdot n_h \geq 0 \) must be included because of the definition of \( \mathcal{L}_h^F \). Then, defining the vectors \( v_h = u_h \) and \( p_h = \delta h + t(\vartheta_h - \delta h) \in \mathcal{X}_h \) with \( t \in (0, 1) \), and taking a finite set of collocation points \( c \in \Gamma_h \).

\[
\mathcal{L}_h^F(\delta_h + t(\vartheta_h - \delta_h)) - \mathcal{L}_h^F(\delta_h) + \lambda_h \cdot [-t(\vartheta_h - \delta_h)] + \frac{\gamma}{2} [t(\vartheta_h - \delta_h)]^2 \geq 0, \quad \forall c \in \Gamma_h
\]

Dividing by \( t \) and taking the limit when \( t \to 0^+ \) the first term becomes a directional derivative of \( \mathcal{L}_h^F \) (if it is defined) with direction \( \vartheta_h - \delta_h \):

\[
(\partial_h \mathcal{L}_h^F - \lambda_h) \cdot (\vartheta_h - \delta_h) \geq 0, \quad \forall c \in \Gamma_h, \forall \vartheta_h \in \mathcal{X}_h
\]

So the last equation in 29 holds. It can be seen that \( (J_h u_h) \cdot n_h \geq 0 \) with \( u_h \in \mathcal{K}_h \subset \mathcal{K} \) is a minimum of the discrete functional \( \mathcal{L}_{u_h} \).
Now, the proof of the uniqueness of the saddle point of $L_h$ is presented. Let's suppose the existence of two saddle points $(u_h, \delta_h, \lambda_h)$ and $(u_h^*, \delta_h^*, \lambda_h^*) \in \mathcal{V}_h \times \mathcal{X}_h \times \mathcal{W}_h$ of $L_h : \mathcal{V}_h \times \mathcal{X}_h \times \mathcal{W}_h \rightarrow \mathbb{R}$. Since they correspond to saddle points, $u_h \in \mathcal{K}_h$ and $u_h^* \in \mathcal{K}_h$ are both solutions of the direct problem, and so $u_h = u_h^*$ because of the uniqueness of the solution. From equation 29, it can be concluded that $\delta_h = \delta_h^*$. Finally, the equation 28 makes $\lambda_h = \lambda_h^*$ if the proposed space discretization fulfill the inf-sup condition 25.

To prove the existence of the saddle point of $L_h$, the unique minimum $u_h \in \mathcal{K}_h$ of functional $L_{uh}$ is considered. By the reciprocity of the Brezzi theorem [4], a unique $\lambda_h \in \mathcal{W}_h$ exists if condition 25 is fulfilled, with the following restriction:

$$\int_{\Gamma_h} \lambda_h \cdot (J_h v_h) \ d\Gamma_h \leq 0, \ \forall v_h \in \mathcal{V}_h$$

The proof of the double inequality 26 can be completed following the same way used to prove proposition 2.2.

\[ \square \]

**Remark 2.1.** Observe that by relation 29, the equations 27 y 28 are equivalent to the stationary point stated in 11.

**Remark 2.2.** The discretization used for the numerical examples presented below, correspond to $\mathcal{V}_h = P^2_c$ and $\mathcal{W}_h = P^1_c$. For $\mathcal{W}_h = P^2_c$, the inf-sup condition is also satisfied, but this discretization is computationally more expensive.

**Remark 2.3.** Doyen [5], shows that the distance from the approximate solution to the exact one, fulfils the following inequality:

$$\| u - u_h \|_\mathcal{V} + \alpha h^{1/2} \| \lambda - \lambda_h \|_\mathcal{W} \leq Ch^{1+\nu} \tag{31}$$

with $C > 0$ and $0 < \nu < \frac{1}{2}$ which means that the solution of the approximate problem is bounded and then, it is well-posed.

### 3 SOME TOPICS ABOUT THE COHESIVE POTENTIAL

The adaptation of the cohesive zone model based on a simplification of the model proposed by Park et al. [10] in terms of an equivalent gap $\delta_{eq} = \sqrt{\delta \cdot \delta}$, is described in this section.

The potential expressed in equation 5 is defined as follows:

$$\psi(\delta_{eq}) = \begin{cases} G_c \left( 1 - \left( \frac{\delta_{eq}}{\delta_c} \right)^{\beta} \right), & \text{if } \delta_{eq} \leq \delta_c \\ G_c, & \text{if } \delta_{eq} > \delta_c \end{cases} \tag{32}$$

The cohesive energy is defined with the expression $G_c = \frac{1}{\beta} \sigma_c \delta_c$, where $\sigma_c$ is the crack initiation stress or critical stress, $\delta_c$ is the critical displacement and $\beta$ is a parameter that defines the shape of the cohesive law. The sub gradients in each phase, are detailed below.
Adhesion: This branch is activated when $\delta = 0$ and $\kappa = 0$.

$$\partial_\delta \Psi (\delta, \kappa) = \left\{ t \in H^{-\frac{1}{2}} \mid \| t \|_+ \leq \sigma_c \right\} \quad (33)$$

Unloading: is activated when $\delta_{eq} \leq \kappa$ with $\kappa > 0$, so the sub gradient is:

$$\partial_\delta \Psi (\delta, \kappa) = \left\{ t_n \mathbf{n} + \rho \delta \mid \rho = \frac{\psi' (\kappa)}{\kappa} \land t_n \leq 0 \land \delta_n \geq 0 \land t_n \delta_n = 0 \right\} \quad (34)$$

Damage: If $\delta_{eq} > \kappa$, the crack propagate over the interface. The sub gradient follows:

$$\partial_\delta \Psi (\delta, \kappa) = \left\{ t_n \mathbf{n} + \delta \frac{\psi' (\delta_{eq})}{\delta_{eq}} \mid t_n \leq 0 \land \delta_n \geq 0 \land t_n \delta_n = 0 \right\} \quad (35)$$

where $\kappa$ represents a memory variable equals to the max $\delta_{eq}$ reached in previous steps. In the previous equations, $t_n$ represents the normal projection of $t$, and:

$$\| t \|_+ = (\langle t \rangle_+ \cdot \langle t \rangle_+)^{\frac{1}{2}}, \quad \langle t \rangle_+ = \langle t_n \rangle \mathbf{n} + t_\parallel$$

being $\langle \bullet \rangle$ the Macaulay brackets, and $t_\parallel$ the tangential projection of $t$.

4 NUMERICAL EXAMPLE

A numerical example consisting of a hard inclusion debonding in a soft matrix is presented. The cell is subject to a prescribed displacement on the top of it. A path-following technique based on the released energy control to get the snap-back instability is used. Figure 2 shows the input data of the model.

![Figure 2](image)

_A parametric study varying the cohesive energy is performed. The left figure in 3 shows the mechanical response of the cell in terms of the prescribed displacement and the force..._
in the top edge, making evident the unstable propagation of the fracture for cohesive energies lower than 0.002. When all the energy in the interface is released and the crack become self-closed all responses resemble the response of the same cell but with a pore or void instead of an inclusion (indicated with dot line). The prescribed displacement versus the normal jump displacement or normal gap in point P is plotted in figure 3, showing that the crack opens no matter if the top displacement load increases or decreases, which means that for unstable behavior, the crack propagation is due the displacement history but not the actual one.

5 CONCLUSIONS

An Augmented Lagrangian formulation with special concern about the stability of the solution was studied. The equivalence between the direct and mixed problem was proved, both in continuous and discrete spaces. The second was equipped by an a-priori error estimate in terms of the mesh size $h$.

The existence and uniqueness of the solution was proved like an extension of the theorems for contact mechanics presented by Chen [6].

A non linear cohesive zone model with linear unloading was implemented in the presented mixed variational framework. The debonding phenomenon of a stiff particle immersed in a soft matrix was simulated evidencing that, for lowest fracture energies, an unstable crack propagation occurs. Once the crack starts its propagation in unstable regime, the normal gap increases despite the decrease of the prescribed displacement.

REFERENCES


MULTI-REGION BOUNDARY ELEMENT METHOD AND TANGENT OPERATOR TECHNIQUE APPLIED TO CRACK PROPAGATION MODELLING IN CONCRETE STRUCTURES

SÉRGIO G.F. CORDEIRO¹, EDSON D. LEONEL²

¹University of São Paulo, Department of Structural Engineering
Av. Trabalhador São Carlense, 400, Centro, CEP: 13.566-590 São Carlos-SP, Brazil
sergio.cordeiro@usp.br

² University of São Paulo, Department of Structural Engineering
Av. Trabalhador São Carlense, 400, Centro, CEP: 13.566-590 São Carlos-SP, Brazil
edleonel@sc.usp.br

Key words: Multi-Region BEM, Cohesive Crack, Concrete Structures, Tangent Operator.

Abstract. The cohesive crack model has been applied by the scientific community and it has already demonstrated its efficiency and accuracy in simulating crack growth phenomena in concrete bodies. In addition to that, the Boundary Element Method (BEM) is recognized as a robust and efficient numerical technique capable to address fracture problems. In this regard, the present study aims at coupling the cohesive crack model with a multi-region BEM formulation to simulate the mechanical behaviour of two-dimensional concrete structure. The cohesive cracks are modelled along the interfaces of multi-region domains. It allows the representation of the structural nonlinear mechanical behaviour as well as the degradation process at the interfaces. As the surface tractions and crack opening displacements at fractured interfaces are interdependent, this problem is solved in the context of nonlinear solutions. The nonlinear system of equations is solved using a Tangent Operator (TO) scheme, where tangent prevision and tangent correction steps are required. Such scheme has demonstrated to assure better convergence and accuracy than the classical Newton Raphson approach as the derivative set of algebraic equations incorporates the cohesive laws. The determination of the tangent operator terms for the cohesive laws is the main contribution of this study. To validate the proposed formulation, two applications concerning concrete beams in fracture tests were performed. The results achieved were compared with numerical and experimental responses available in literature. Apart from the good agreement among the obtained results, faster convergence was verified using TO instead of the classical scheme.
1 INTRODUCTION

The collapse process of engineering structures occurs by the growth of internal discontinuities (cracks) which lead to the structural failure by fracture. Among classical engineering materials, concrete is commonly used around the world due to its production cost, geometry adaptability, chemical properties and strength capacities. Therefore, concepts of fracture mechanics applied to concrete have been widely discussed and studied by the scientific community in recent years. In such type of material, the fracture phenomenon causes nonlinear structural response, due to the mechanical degradation processes. Consequently, nonlinear models were proposed along the past years to represent the concrete fracture [1-3]. One interesting alternative to handle the crack growth in quasi-brittle materials, such as the concrete, is the fictitious crack model or cohesive crack model. This approach was introduced by [1] and extensively adopted in numerical applications of fracture problems [4-7]. Based on cohesive approach, crack propagation modelling of quasi-brittle materials requires the solution of a nonlinear problem, generally relating the Crack Opening Displacements (COD) to tractions along the crack surfaces.

In this regard, this paper addresses the mechanical analysis of crack growth in concrete beams using a nonlinear Boundary Element Method (BEM) formulation and an efficient algorithm to solve the nonlinear system of equations based on the Tangent Operator (TO) scheme. The BEM is a robust numerical technique for modelling crack surfaces as only boundary surfaces discretization is required [8]. Therefore, this numerical technique is well adapted to solve fracture mechanics problems. Moreover, the stresses singularities present at the crack tip are accurately represented due to the use of fundamental solutions. The formulation applied in this study involves the sub-region technique and the cohesive model. This first technique divides the entire domain into sub-regions in which, for elastic problems, compatibility of displacements and equilibrium of forces are enforced along the interfaces that connect them. In cohesive crack problems, the displacement compatibility is no longer verified. Therefore, cohesive laws are required to relate the COD with the tractions at crack surfaces. As a result, cracks in the present work are assumed to grow along sub-region interfaces. One advantage of this approach concerns the possibility to analyse crack propagation in composed nonhomogeneous materials [9].

The nonlinear problem is solved using the TO approach. This procedure is derived as it assures better convergence and accuracy than the classical Newton-Raphson approach. This procedure achieves the solution using an initial try and corrections steps that account the tangent direction of the nonlinear response. This tangent search is obtained as the derivative set of algebraic equations is constructed considering the analytical function of the cohesive laws. This type of operator has already been successfully used in the literature for dealing many different nonlinear engineering problems. For instance, [10], [11] and [12] applied this operator on localization phenomenon, cohesive crack growth, and contact problems, respectively.

Two applications were performed with the developed formulation and the results were compared with experimental and numerical responses available in literature to prove its robustness and accuracy. It was also verified that the TO provides faster convergence and computational time consuming for the nonlinear solution when compared with classical Newton method.
2 BOUNDARY INTEGRAL EQUATIONS

The BEM has been widely applied in engineering fields such as contact problems, fatigue and fracture mechanics due to its high precision and robustness in modelling strong stress concentration. Considering a two-dimensional homogeneous elastic domain, $\Omega$, with boundary, $\Gamma$, the equilibrium equation, written in terms of displacements, is given by:

$$u_{i,j} + \frac{1}{1-2\nu}u_{j,i} + \frac{b_i}{\mu} = 0$$

(1)

where $\mu$ is the shear modulus, $\nu$ is the Poisson’s ratio, $u_i$ are components of the displacement field, and $b_i$ are body forces. Using Betti’s theorem, the singular integral representation, written in terms of displacements, is obtained, disregarding body forces, as follows:

$$c_{ij}(s,f)u_j(s) + \int_{\Gamma} P_{ij}^*(s,f)u_j(f)\,d\Gamma = \int_{\Gamma} U_{ij}^*(s,f)p_j(f)\,d\Gamma$$

(2)

where $p_j$ and $u_j$ are tractions and displacements on the boundary respectively, the free term $c_{ij}$ is equal to $\delta_{ij}/2$ for smooth boundaries and $P_{ij}^*$ and $U_{ij}^*$ are the fundamental solutions for tractions and displacements written for the source point $s$ [13].

Differentiating (2) with respect to the directions $x, y$ an integral equation written in terms of strains is obtained. Applying, for instance, the generalized Hooke’s law the integral representation of stresses for a boundary source point is obtained by carrying out the relevant limits. Then, the Cauchy formula may be applied to obtain the hyper-singular integral equation which is written in terms of tractions as follows:

$$\frac{1}{2}p_j(s) + \eta_k \int_{\Gamma} S_{kij}(s,f)u_k(f)\,d\Gamma = \eta_k \int_{\Gamma} D_{kij}(s,f)p_k(f)\,d\Gamma$$

(3)

$S_{kij}$ and $D_{kij}$ contain the new kernels computed from $P_{ij}^*$ and $U_{ij}^*$ respectively [13] and $\eta_k$ is the outward normal vector.

Applying separately Equations (2) or (3) it is possible to construct the algebraic system of equations for two-dimensional non-cracked domains. However, the single use of just one integral equation for modelling crack surfaces lead to a singular matrix when assembling the BEM system. Therefore, to analyse solids containing cracks, some alternatives can be found in literature. One approach is modelling crack surfaces along the interfaces of multi-region problems, as performed in this study. Other alternatives such as the Dual-BEM [13] that use simultaneously both integral equations or even the dipole formulation [11] may be applied.

To assemble the system of BEM equation, as usual, Equation (2) or (3) is transformed into algebraic relations by discretizing the boundary and interfaces into elements along which displacements and tractions are approximated. Besides that, one has to select a convenient number of collocation points to obtain the algebraic representations. The algebraic equations for boundary nodes are calculated considering the boundary collocation points.

After determining the displacement and tractions fields at the boundary, internal values for displacements, stresses and strains can be achieved. Internal displacements are determined using the integral Equation (2) with the source point $s$ located at the interior of the domains. In this case, the free term $c_{ij}$ becomes $\delta_{ij}$. On the other hand, the stress field at internal nodes is obtained through the integral representation of stresses [8].
3 MULTI-REGION BEM AND COHESIVE CRACKS

As presented at the introduction, an alternative for modelling crack surfaces is by the multi-region technique. In this approach, the crack surfaces are discretized with oppositely oriented interface elements. To analyse quasi-brittle fracture in concrete, the cohesive crack model was incorporated to these interface elements introducing nonlinear behaviour for the structures. Some details of this formulation and also the TO algorithm to solve the nonlinear system are following presented.

3.1 Cohesive cracks along multi-region interfaces

Considering multi-region technique, the classical set of BEM algebraic equations is written for a given domain accounting the location of the collocation points, i.e. source points. These points may be located at the interface opposite elements at left \((L)\) or right \((R)\) sides, as illustrate in Figure 1, or even at the boundary \((B)\) of the problem.

\[
H_{ij}^{B}u_{j}^{B} + \sum_{k=1}^{NI}(H_{ij}^{L(k)}u_{j}^{L(k)} + H_{ij}^{R(k)}u_{j}^{R(k)}) = G_{ij}^{B}p_{j}^{B} + \sum_{k=1}^{NI}(G_{ij}^{L(k)}p_{j}^{L(k)} + G_{ij}^{R(k)}p_{j}^{R(k)}) \tag{4}
\]

As illustrated in Figure 1, displacements and tractions must be rotated from \(x, y\) global orientation to normal \((\eta)\) and tangent \((t)\) directions of each crack side. Defining \(u_{j}^{S(k)}, u_{j}^{T(k)}, p_{j}^{S(k)}, p_{j}^{T(k)}\) as the normal and tangent components of displacement and tractions at a given side \(S\) of an interface \(k\), respectively, and imposing the boundary conditions it is possible to rewrite the Equation (4) as follows:

\[
A_{ij}^{B}u_{j}^{B} + \sum_{k=1}^{NI}
\begin{bmatrix}
H_{ij}^{L(k)}u_{j}^{L(k)} + H_{ij}^{T(k)}u_{j}^{T(k)} \\
H_{ij}^{R(k)}u_{j}^{R(k)} + H_{ij}^{T(k)}u_{j}^{T(k)}
\end{bmatrix}
= F_{i}^{B} + \sum_{k=1}^{NI}
\begin{bmatrix}
G_{ij}^{L(k)}p_{j}^{L(k)} + G_{ij}^{T(k)}p_{j}^{T(k)} \\
G_{ij}^{R(k)}p_{j}^{R(k)} + G_{ij}^{T(k)}p_{j}^{T(k)}
\end{bmatrix} \tag{5}
\]

**Figure 1:** Tractions and displacements rotation at opposite interface elements
In which, $H^n_{ij}, H^t_{ij}$ and $G^n_{ij}, G^t_{ij}$ are the collection of terms that multiply the normal and tangent displacements and tractions after the rotation. $x^B_i$ are the boundary unknown values, $A^B_{ij}$ is composed by columns of $H^B_{ij}$ and $G^B_{ij}$ referent to these same values and $F^B_i$ is a vector that computes the contribution of the prescribed quantities at the boundary. To solve such problem it is necessary to impose the following equilibrium condition for the interface tractions:

$$p^\eta_j - p^\eta_j = 0$$
$$p^t - p^t_j = 0$$  \hspace{1cm} (6)

Therefore, using (6), (5) becomes:

$$A^B_{ij} x^B_j + \sum_{k=1}^{N_I} \left[ \begin{array}{c}
H^n_{ij} l^{(k)} u^n_{ij} l^{(k)} + H^t_{ij} l^{(k)} u^t_{ij} l^{(k)} + \varepsilon
G^n_{ij} l^{(k)} + G^t_{ij} l^{(k)} \end{array} \right] = F^B_i + \sum_{k=1}^{N_I} \left[ \begin{array}{c}
(G^n_{ij} l^{(k)} + G^t_{ij} l^{(k)} p^n_j l^{(k)} + G^t_{ij} l^{(k)} p^n_j l^{(k)} \end{array} \right] \hspace{1cm} (7)

At the above system of equations, no condition was imposed for interface displacements. Considering, at first, that the interface surfaces are free to open and slide, the following condition can be written:

$$-u^\eta_j - u^\eta_j = COD_j$$
$$u^t_j + u^t_j = CSD_j$$  \hspace{1cm} (8)

where $COD_j$ are the Crack Opening Displacements and $CSD_j$ the Crack Sliding Displacements on opposite interface collocation points. Including Condition (8) at (7) and assuming $COD_j = CSD_j = 0$ it is possible to solve the problem for an elastic interface, Figure 2 (left).

Instead of elastic conditions, the cohesive crack model can also be incorporated at the interfaces as showed in Figure 2 (right). For that, the tensile strength $f_t$ and a softening cohesive law for the material must be known. Among the proposed cohesive laws for representing concrete cracking, three are the most commonly achieved in literature and are...
represented as linear, bilinear and exponential curves [11]. Then, these laws were adopted in this study for the applications. When the normal tractions $p^n_j$ at some interface nodes overcome the value of $f_t$, the previously calculated elastic tractions must be corrected following the presented cohesive condition:

$$p^n_j = p^c(COD) + \Delta p^n_j^{exc}$$

(9)

Condition (9) indicates that the normal tractions $p^n_j$ are function of the normal traction excess $\Delta p^n_j^{exc}$, that must be taken to zero in an iterative procedure, and of the crack opening that is governed by a cohesive law $p^c(COD)$. This traction excess is computed as the difference between elastic and cohesive tractions, defining the non-equilibrated stress vector.

The second condition in (9) appoints for a brittle shear behaviour, which means $p^n_j = 0$ after the elastic limit. Figure 2 (b) indicates that after the opening $COD$ overcome the critical value, $\Delta u^{cr}$, normal cohesive tractions are also assumed as zero. Hence, to consider cohesive behaviour, Condition (9) must be incorporate at the system of BEM equations resulting in:

$$A_{ij}^B x_j^B + \sum_{k=1}^{N_l} \left[ (H^n_{ij}^{l(k)} - H^n_{ij}^{r(k)}) u^n_j^{l(k)} - H^n_{ij}^{r(k)} COD_{ij}^{(k)} + (H^t_{ij}^{l(k)} - H^t_{ij}^{r(k)}) u^t_j^{l(k)} + H^t_{ij}^{r(k)} CSD_{ij}^{(k)} \right] = F_i^B + \sum_{k=1}^{N_l} \left( (G^n_{ij}^{l(k)} + G^n_{ij}^{r(k)}) (p^c(COD_{ij}^{(k)}) + \Delta p^n_j^{exc(k)}) \right)$$

(10)

Once the cohesive law was included, System (10) becomes nonlinear due to the interdependency of the variables $p^c$ and $COD$. To solve such nonlinear system, the TO approach is introduced in the next topic.

### 3.2 Nonlinear solution using Tangent Operator approach

In order to solve the nonlinear problem, it is important to rewrite the System (10) as a function of its variables as follows:

$$Y(x_j^B, u^n_j^{l(k)}, COD_{ij}^{(k)}, u^t_j^{l(k)}, CSD_{ij}^{(k)}, F_i^B) =$$

$$A_{ij}^B x_j^B + \sum_{k=1}^{N_l} \left( (H^n_{ij}^{l(k)} - H^n_{ij}^{r(k)}) u^n_j^{l(k)} - \sum_{k=1}^{N_l} H^n_{ij}^{r(k)} COD_{ij}^{(k)} \right) + \sum_{k=1}^{N_l} \left( (H^t_{ij}^{l(k)} - H^t_{ij}^{r(k)}) u^t_j^{l(k)} + \sum_{k=1}^{N_l} H^t_{ij}^{r(k)} CSD_{ij}^{(k)} \right)$$

$$- \sum_{k=1}^{N_l} \left( G^n_{ij}^{l(k)} + G^n_{ij}^{r(k)} p^c(COD_{ij}^{(k)}) - F_i^B \right) = 0$$

(11)

In (11), traction excess $\Delta p^n_j^{exc(k)}$ was collected at the vector $F_i^B$ once such values are prescribed as traction corrections at the cohesive surfaces. To solve the nonlinear system by
TO scheme, Equation (11) must be differentiated with respect to the variables.

\[
\frac{\partial Y}{\partial \text{variables}} = \frac{\partial Y}{\partial x^B} + \frac{\partial Y}{\partial u_j^{l_j}} + \frac{\partial Y}{\partial COD_k}^B + \frac{\partial Y}{\partial \Gamma}^B = 0
\]  

(12)

Proceeding the differentiation, the derivative set of BEM algebraic equation can be presented as follows:

\[
\begin{align*}
A_{ij}^B \partial x_j^B + \sum_{k=1}^{N_l} (H_{ij}^{\eta_j} - H_{ij}^{R_j}) \partial u_j^{l_j} &- \sum_{k=1}^{N_l} H_{ij}^{R_j} \partial COD_k^B \\
\sum_{k=1}^{N_l} (H_{ij}^{\Gamma_j} - H_{ij}^{R_j}) \partial u_j^{l_j} + \sum_{k=1}^{N_l} H_{ij}^{R_j} \partial COD_k^B &- \sum_{k=1}^{N_l} (G_{ij}^{\eta_j} + G_{ij}^{R_j}) \frac{\partial p^C}{\partial COD_k^B} \partial COD_k^B = \partial F_i^B
\end{align*}
\]  

(13)

In (13), the cohesive tractions \( p^C \) are differentiated with respect to the openings \( COD_j \). Therefore, the analytical expression of a cohesive law is required. As the nonlinear problem is solved incrementally, System (13) can be expressed in terms of increments as follows:

\[
\begin{align*}
\Delta x_j^B & A_{ij}^B, \\
\Delta u_j^{l_j} & (H_{ij}^{\eta_j} - H_{ij}^{R_j}), \\
\Delta COD_k^B & \sum_{k=1}^{N_l} (-H_{ij}^{R_j} - (G_{ij}^{\eta_j} + G_{ij}^{R_j}) \frac{\partial p^C}{\partial COD_k^B}) \partial COD_k^B, \\
\Delta \Gamma_j^{l_j} & \sum_{k=1}^{N_l} (H_{ij}^{\Gamma_j} - H_{ij}^{R_j}) \\
\Delta \Gamma_j^{R_j} & \sum_{k=1}^{N_l} H_{ij}^{R_j}
\end{align*}
\]  

\[
\begin{bmatrix}
\Delta x_j^B \\
\Delta u_j^{l_j} \\
\Delta COD_k^B \\
\Delta \Gamma_j^{l_j} \\
\Delta \Gamma_j^{R_j}
\end{bmatrix} = A_{ij}^B \begin{bmatrix}
A_{ij}^B, \\
(H_{ij}^{\eta_j} - H_{ij}^{R_j}), \\
\sum_{k=1}^{N_l} (-H_{ij}^{R_j} - (G_{ij}^{\eta_j} + G_{ij}^{R_j}) \frac{\partial p^C}{\partial COD_k^B}) \partial COD_k^B, \\
\sum_{k=1}^{N_l} (H_{ij}^{\Gamma_j} - H_{ij}^{R_j}) \\
\sum_{k=1}^{N_l} H_{ij}^{R_j}
\end{bmatrix}^{-1}
\]  

(14)

At this incremental system, the terms \( \partial p^C / \partial COD_k^B \) are the so called Tangent Operator. These terms take information about the tangent of the cohesive laws. Therefore, when incorporated to the set of BEM equations makes the search for the nonlinear solution be performed along the tangent direction. The nonlinear solution of (14) is obtained in an iterative process. Hence, the traction excesses \( \Delta p^j \text{exc} \) of a previous iteration and the corresponding tangent derivatives \( \partial p^C / \partial COD_k^B \) are accounted in (14) to find the next iteration solution. The adopted stop criterion for such iterative procedure was based on the norm of the vector \( \Delta p^j \text{exc} \), non-equilibrated stress vector, which must be smaller than a prescribed tolerance. As the boundary load is prescribed incrementally, such iterative procedure must be done at each incremental step. By the end of a converged load step, interface regions that were assumed into elastic regime must be checked again to assure that none of the nodes the tractions exceed the tensile strength \( f_t \).
4 APPLICATIONS

Two applications of fracture tests in concrete beams are presented. The first concerns a three-point bended beam test and the second a four-point mixed mode cohesive crack growth test.

4.1 Three-point bended beam with single notch

The first application concerns a mode I fracture test analysed experimentally by [5]. The specimen consists of three point bended concrete beam with an initial notch. This structure was modelled numerically by [11] with an alternative BEM formulation. To describe the three cohesive laws, two parameters were considered. The tensile strength $f_t = 3\, MPa$ and the fracture energy $G_f = 75\, N/m$. Figure 3 illustrates the problem presenting the geometrical data and the material elastic proprieties.

![Figure 3: Three point bended beam test](image)

The problem was divided into two sub-regions by a vertical interface where the main crack propagates in mode I. The adopted boundary mesh is composed by 55 cubic and 29 quadratic discontinuous elements with 307 source points. The prescribed displacements were imposed into 100 incremental steps and the considered tolerance for the stop criterion was $10^{-6} \, MPa$. Figure 4 shows the initial and deformed configuration of the structure with displacements magnified by 100.

![Figure 4: initial mesh and 100 times deformed mesh](image)

Displacement fields in the domains interior and boundary are presented in Figure 5 by colour scale graphics which show the discontinuous crack behaviour.
During the analysis, interface tractions were monitored and its cohesive behaviour is illustrated in Figure 6 (left) for linear, bilinear and exponential cohesive laws. The nonlinear structural behaviour was also monitored through load-displacement curves. The considered load was the equivalent force at the superior beam surface. The displacement values were taken as the average between vertical displacements of two points at the inferior beam face closest to the notch and one at each fracture side. Such nonlinear responses are illustrated in Figure 6 (right) were experimental and numerical curves obtained by previous references are also presented.

The load-displacement curves show that the developed formulation is capable to represent the nonlinear mechanical structural behaviour. Besides, for the three cohesive laws, the values of load pic remained close to the experimental one. The softening part of the obtained curves reproduced quite well the experimental results. On the other hand, the up part of such curves presented a rigid behaviour than the experimental one.
In order to illustrate how the classical Newton method differs from the TO scheme with respect to the nonlinear solution search, the load-displacement curve for the linear law was again plotted. However, in this time, only the results for multiples of 10 steps were presented. These results for both approaches are presented in Figure 7.

Concerning these graphics, the faster convergence of TO is easily understood due to its search being performed at the tangent direction of the solution curve. To quantify the efficiency of the TO approach and classical Newton Raphson scheme, the number of iteration performed during the analyses were monitored and presented in Table 1. The efficiency improvement in terms of iteration number reduction obtained with the TO were remarkable. Such reduction reaches, for the best case, 92% and 77%, for the worst case.

![Figure 7: Nonlinear solution search: Classical Newton-Raphson versus Tangent Operator](image)

<table>
<thead>
<tr>
<th>Cohesive law</th>
<th>Classical Scheme</th>
<th>Tangent Operator</th>
<th>reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>2150</td>
<td>180</td>
<td>~92%</td>
</tr>
<tr>
<td>Bilinear</td>
<td>2157</td>
<td>199</td>
<td>~91%</td>
</tr>
<tr>
<td>Exponential</td>
<td>2067</td>
<td>468</td>
<td>~77%</td>
</tr>
</tbody>
</table>

### 4.2 Four-point mixed mode crack growth test. Shear beam with double notch

The second application concerns the mechanical analysis of a multiple mixed mode cohesive crack growth case. The specimen is composed by a four-point shear beam test analysed numerically by [6] with the Finite Element Method (FEM) and by [4] with the BEM. Both authors adopted the linear cohesive law. The crack paths determined in these works were adopted as interfaces in the model proposed in this work. Therefore, the entire domain was splitted into three sub-regions as presented in Figure 8, which also shows geometrical data and material elastic properties considered. The crack behaviour was represented using linear,
bilinear and exponential cohesive laws. The adopted tensile strength $f_t$ was defined as $2\,MPa$ and the fracture energy $G_f$ was assumed equal to $100\,N/m$.

The adopted boundary mesh was composed by 72 cubic and 30 quadratic discontinuous elements with 378 source points. The prescribed displacements and stop criterion tolerance were the same of the previous application. Figure 9 illustrates the initial and deformed configuration of the structure with displacements magnified by 500.

![Four point shear beam test](image)

**Figure 8:** Four point shear beam test

![Initial mesh and 500 times deformed mesh](image)

**Figure 9:** Initial mesh and 500 times deformed mesh

Displacement fields into the domains interior and boundary are presented in Figure 10, where the colour scale graph shows the discontinuity introduced by the crack.

![Color scale displacement field](image)

**Figure 10:** Color scale displacement field

Once again, traction cohesive responses and load-displacement curves were monitored during the analysis. The load was considered as the equivalent force at the beam superior face. However, the displacements values were considered as the prescribed ones. Figure 11 (left) shows the cohesive traction responses and Figure 11 (right) the load-displacement curves. Good agreement is observed among the numerical results considered.
The load-displacement curves show that the developed formulation was capable to represent the nonlinear structural behaviour for this application. As authors [6] and [4] used the linear cohesive law, the results obtained with such law using the developed technique presented the best agreement. The three cohesive laws applied gave a good accuracy in terms of maximum resistant load.

The efficiency in terms of iteration number is presented in Table 2, where the number of iterations required by TO is compared with the classical approach.

![Figure 11: Cohesive tractions at the interface (left) and nonlinear structural answer (right)](image)

<table>
<thead>
<tr>
<th>Cohesive law</th>
<th>Classical Scheme</th>
<th>Tangent Operator</th>
<th>reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>1281</td>
<td>100</td>
<td>~92%</td>
</tr>
<tr>
<td>Bilinear</td>
<td>3346</td>
<td>220</td>
<td>~93%</td>
</tr>
<tr>
<td>Exponential</td>
<td>3527</td>
<td>579</td>
<td>~84%</td>
</tr>
</tbody>
</table>

5 CONCLUSIONS

- A BEM formulation has been presented for modelling fracture test in concrete structures using the cohesive crack model. For the considered applications, the results obtained showed good agreement with numerical and experimental results available in literature.

- Furthermore, the TO scheme has demonstrated its efficiency as it required a number of iterations considerably lesser than the classical Newton-Raphson approach. Therefore, such operator may lead to remarkable gains in terms of computational cost when applied in most complex engineering problems involving failure by fracture.
ACKNOWLEDGEMENTS

Sponsorship of this research project by the São Paulo Research Foundation (FAPESP), grant number 2014/18928-2, is greatly appreciated.

REFERENCES


FREE VIBRATION OF TIMOSHENKO MICRO BEAMS USING THE MODIFIED COUPLE STRESS THEORY

GRACIELA I. GUERRERO¹², DIANA V. BAMBILL²³* AND DANIEL H. FELIX²
¹Facultad de Ingeniería, Universidad Nacional de La Patagonia San Juan Bosco
km 4, 9000 Comodoro Rivadavia, Argentina
graisague@hotmail.es, http://www.unp.edu.ar/gmc

²Departamento de Ingeniería, Universidad Nacional del Sur
Alem 1253,8000 Bahía Blanca, Argentina
http://www.uns.edu.ar

³ Consejo Nacional de Investigaciones Científicas y Técnicas, Alem 1253, Bahía Blanca,
8000, Argentina
http://www.conicet.gov.ar

Key words: Free Vibrations, micro beam, Timoshenko, Elastic Support, MCST, Ritz Method.

Abstract. A free vibration analysis of elastically restrained micro beams is performed based on Timoshenko beam theory. The analytical formulation is obtained applying the modified couple stress theory (MCST). The elastic boundary conditions are materialized by rotational and transnational classical springs at both ends of the beam. Poisson’s effect on free vibrations is analyzed in conjunction with the material length scale. A study relating the classical elasticity model and the couple stress strain model is also presented. Numerical results for various boundary conditions are determined using the Ritz method, and they are compared wherever is possible with published results.

1 INTRODUCTION

Yang and co-workers, (2002) [1] analyzed the deformation behavior of materials in the micron scale by the couple stress theory formulation. They proposed an additional equilibrium relation to govern the behavior of the couples and developed a linear elastic model for isotropic materials. The model involves, besides the Lamé’s constants, only one additional material constant [2]. Recently Araújo dos Santos and Reddy [2] present a study of vibration of vibration of Timoshenko beams using non-classical elasticity theories, they compared classical elasticity, nonlocal elasticity, and modified couple stress theories for free vibration analysis of Timoshenko beams. [3] presented a comparison among classical elasticity, nonlocal elasticity, and modified couple stress theories for free vibration analysis of Timoshenko beams and studied the influence of rotary inertia and nonlocal parameters on natural frequencies. Ma et al. [4] developed a microstructure-dependent Timoshenko beam model based on a modified couple stress theory and Hamilton’s principle. Akgöz and Civalek [5] developed size-dependent higher-order shear deformation beam model which captures both the microstructural and observed that effect of shear deformation becomes more significant for both smaller slenderness ratios and higher modes shear deformation effects without the need for any shear correction factors, they observed that effect of shear deformation becomes more significant for smaller slenderness ratios and higher modes. Park and Gao [6] provided a variational formulation for the modified couple stress theory of Yang

* Corresponding author
et al. by using the principle of minimum total potential energy. They solved a simple shear problem analytically and determined that the length scale parameter can have a significant effect on material responses. Reddy (2007) [7] reformulated the Euler–Bernoulli, Timoshenko, Reddy, and Levinson beam theories using the nonlocal differential constitutive relations of Eringen. Reddy (2011) [8] reformulated classical and shear deformation beam and plate theories using the nonlocal differential constitutive relations of Eringen and the von Kármán nonlinear strains. Wang et al. (2010) [9] developed a micro scale Timoshenko beam model developed based on strain gradient elasticity theory. They assumed that the model can degenerate into the modified couple stress Timoshenko beam model or even the classical Timoshenko beam model if two or all material length scale parameters are taken to be zero.

Xia et al. [10] analyzed nonlinear micro beams and investigated the static bending, postbuckling and free vibration. They concluded that indicate that the nonlinearity has a great effect on the static and dynamic behaviors of microscale beams.

In the present study, the free vibration of elastically supported micro beams is analyzed by using a Timoshenko beam model based on the couple stress theory. Not many previous results of free vibration of micro beams with elastic boundary conditions are present in the open literature. Rinaldi et al. (2008) [11] presented a testing method which helps to identify the influence of microfabrication to the support boundary conditions of micro beams and presented results on the characterization of non-classical support boundary conditions of microcantilevers through electro-mechanical testing. Abadyan et al. [12] investigated of elastic boundary condition on the Pull-in instability of beam-type NEMS under van der Waals Attraction. And [13-14] studied elastic boundary conditions and size dependency on the pull-in instability of beam-type NEMS.

In the numerical examples, the first three frequency coefficients of Timoshenko micro beams are presented for various values of rigidity of translational and rotational springs. The accuracy of the results is established by comparison with previously published results for the free vibration analysis of the Timoshenko micro beams with classical boundary conditions.

2 FORMULATION

Figure 1 shows a micro beam with length $L$ and thickness $h$, that has non-classical support boundary conditions. Based on the modified couple stress theory the strain energy can be calculated by [15]

$$U_t = \frac{1}{2} \int \left( \sigma_{ij} \varepsilon_{ij} + m_{ij} \chi_{ij} \right) d\Omega.$$  \hspace{1cm} (1)

where

- $\Omega$: volume of the micro beam
- $\sigma_{ij}$: stress tensor
- $\varepsilon_{ij}$: strain tensor
- $m_{ij}$: deviatoric part of the couple stress tensor
- $\chi_{ij}$: symmetric curvature tensor
- $\lambda = \frac{E}{(1+v)(1-2v)}$: Lamé’s constant
- $\mu = \frac{E}{2(1+v)}$: Lamé’s constant (it is also known as shear modulus $G$)
- $\theta_i = \frac{1}{2} e_{ijk} u_{k,i}$: rotation vector
- $E$: Young’s modulus
- $\nu$: Poisson’s ratio
- $e_{ijk}$: Levi Civita tensor
As it is known, the modified modified couple stress theory is a function of both the strain (conjugated with stress) and the curvature (conjugated with couple stress) and involves, besides the Lamé’s constants, only one additional material constant (a length scale parameter).

The displacement field of any point of the micro beam, according to the Timoshenko beam theory, is defined by:

\[
\begin{align*}
\delta_{ij} & \text{ Kronecker’s Delta} \\
u_i & \text{ displacement field} \\
l & \text{ a material length scale parameter}
\end{align*}
\]

\[u_1 = u(x,t) = z\phi(x,t), \quad (2)\]
\[u_2 = v(x,t) = 0, \quad (3)\]
\[u_3 = w(x,t). \quad (4)\]

where \(u\) and \(w\) are the axial and transverse displacements and \(\phi\) is the angle of bending rotation of the cross section of the beam, \(\text{Figure 2}\) and \(t\) is the time. In the Timoshenko beam theory, plane sections remain plane but are no normal to the longitudinal axis. The difference between the normal to the longitudinal axis and the plane section rotation is the shear deformation \(\gamma\).

\[\phi = \gamma - \frac{\partial w}{\partial x} \quad (5)\]

\[\gamma \quad -\frac{\partial w}{\partial x} \quad \phi\]

\(\text{Figure 1: Micro beam configuration and coordinate system}\)

\(\text{Figure 2: Rotation angle}\)

The expressions of the above-mentioned tensors are in Eqs. (6-10): stress tensor, strain tensor, deviatoric part of the couple stress tensor and symmetric curvature tensor, respectively.
The strain energy of the micro beam can be written as
\[^{(11)}\]
\[\mathcal{U}_i = \frac{1}{2} \int_A \int_0^L (\sigma_{ij} \dot{\varepsilon}_{ij} + 2 \sigma_{ij} \varepsilon_{ij} + 2 m_{ij} \chi_{ij}) dA dx\]
\[= \frac{1}{2} \int_A \int_0^L \left[ (\lambda + 2\mu) \varepsilon^2 (\phi')^2 + \mu \kappa (\phi + w')^2 + l^2 \frac{\rho}{4} (\phi' - w'')^2 \right] dA dx\]
where the volume is \( \Omega = A L \), \( A \) is the area of the cross section, \( \kappa \) is the shear correction factor, \( I = \int z^2 \, dA \) is the moment of inertia and \( (\cdot)' \) indicates partial derivation with respect to \( x \): \( \phi' = \frac{\partial \phi}{\partial x}, w' = \frac{\partial w}{\partial x}, w'' = \frac{\partial^2 w}{\partial x^2} \).

The kinetic energy of the micro beam is given by
\[^{(12)}\]
\[T = \frac{1}{2} \int_A \int_0^L \left( \rho \left( \dot{w}^2 + \dot{z} \phi \right) \right) dA dx = \frac{1}{2} \int_0^L \rho (A \dot{w})^2 + I (\dot{\phi})^2 dx,\]
where \( \rho \) is the mass density and \( (\cdot)' \) indicates partial derivation with respect to time \( t \).

The energy stored in the micro beam due to the elastic boundary conditions is given by
where $K_{t1}$, $K_{t2}$, and $K_{r1}$, $K_{r2}$ are the translational and rotational spring stiffness constants, ($K_{t1}$ and $K_{r1}$ correspond to the springs located at $x = 0$ and $K_{t2}$ and $K_{r2}$ to the springs located at $x = L$) Figure 1.

3 TOTAL ENERGY OF THE SYSTEM

The total energy $\pi$ of the system is

$$\pi = U_1 + U_2 - T.$$  \hfill (14)

Substituting Eqs. (11), (12) and (13) into Eq. (14) the following equation is obtained

$$\pi = \frac{1}{2}\int_0^L \left[ (\lambda + 2\mu)I(\phi')^2 + \mu AK(\phi + w')^2 + \frac{l^2}{4} A(\phi' - w')^2 \right] dx + \frac{1}{2} K_{t1} w(0,t)^2 + \frac{1}{2} K_{t2} w(L,t)^2 + \frac{1}{2} K_{r2} \phi(L,t)^2 + \int_0^L \rho A (\omega)^2 + l^2 \phi^2 \right] dx. \hfill (15)$$

where

$$(\lambda+2\mu)I=EI \frac{(1-\nu)}{(1+\nu)(1-2\nu)} = EIR \hfill (16)$$

In order to perform a vibration analysis a harmonic motion is assumed with a circular natural frequency, $\omega$,

$$w(x, t) = W(x) e^{iat} \hfill (17)$$

$$\phi(x, t) = \varphi(x) e^{iat} \hfill (17)$$

Replacing $w$ and $\varphi$ in the total energy expression $\pi$ Eq. (15)

$$\pi = \frac{e^{iat}}{2} \left\{ \int_0^L \left[ EIR \left( \frac{d\varphi}{dx} \right)^2 + \mu AK \left( \varphi + \frac{dW}{dx} \right)^2 + \frac{l^2}{4} A \left( \frac{d\varphi}{dx} - \frac{d^2W}{dx^2} \right)^2 \right] dx + K_{t1} W(0)^2 + K_{t2} W(L)^2 + K_{r2} \varphi(L)^2 + \int_0^L \rho A (\omega)^2 (W(x))^2 + \int_0^L m_2 (i\omega)^2 (\varphi(x))^2 dx \right\} \hfill (18)$$

where the rotational inertia is written as $m_2 = \rho L$. By applying the principle of minimum energy of the system, it is obtained the variational expression:

$$\delta \pi = 0.$$

4 METODO DE RITZ

In order to solve the free vibration problem Ritz method is employed and a dimensionless coordinate is assumed as:

$$\tilde{x} = \frac{x}{L}. \hfill (19)$$

Then, the dimensionless amplitudes of the generalized displacements can be expressed as:

$$W(x) = \overline{W(\tilde{x})} L, \hfill (20)$$

$$\varphi(x) = \overline{\varphi(\tilde{x})}. \hfill (21)$$
Replacing Eqs. (20) to (21) in Eq. (18)

\[
\pi = \frac{e^{i\omega t}}{2} \left\{ \int_0^L \left[ EIR \left( \frac{d\Phi}{dx} + \frac{d\bar{W}}{dx} \right)^2 + \mu A \left( \frac{d\bar{Phi}}{dx} + \frac{d\bar{W}}{dx} \right)^2 + \frac{I^2 \mu A}{4L} \left( \frac{d^2\Phi}{dx^2} + \frac{d^2\bar{W}}{dx^2} \right)^2 \right] \ dx \right\ }
\]

\[
+ K_1 \left( \bar{W}(0)L \right)^2 + K_2 \left( \bar{W}(L) \right)^2 + K_3 \left( \bar{W}(1)L \right)^2 + K_4 \left( \bar{Phi}(L) \right)^2
\]

\[
+ \int_0^L \rho A(\omega)^2 \left( \bar{W}(x)L \right)^2 \ dx + \int_0^L m_1(\omega)^2 \left( \bar{Phi}(x) \right)^2 \ dx \}
\]

(22)

Multiplying Eq. (22) by \( \frac{2}{e^{i\omega t}} \) the dimensionless expression of the spring rigidities and the non-dimensional frequency parameter result in

\[
k_1 = K_1 \frac{L}{EI}, \quad k_2 = K_2 \frac{L}{EI}, \quad k_3 = K_3 \frac{L}{EI}, \quad k_4 = K_4 \frac{L}{EI}, \quad (\omega)^2 = \rho A \frac{L^4}{EI}(\omega)^2.
\]

(23)

(24)

The generalized displacement functions Eqs. (20) and Eq. (21) are approximated by the sum of expressions of the form

\[
\bar{W}(x) \approx \sum_{i=1}^{N} C_i \psi_i(x),
\]

(25)

\[
\bar{Phi}(x) \approx \sum_{i=1}^{N} B_i \varphi_i(x),
\]

(26)

and their derivatives

\[
\frac{d\bar{W}(x)}{dx} \approx \sum_{i=1}^{N} C_i \frac{d\psi_i(x)}{dx}, \quad \frac{d\bar{Phi}(x)}{dx} \approx \sum_{i=1}^{N} B_i \frac{d\varphi_i(x)}{dx}.
\]

(27)

Replacing Eqs. (23) to Eq. (27) in Eq. (22):

\[
\frac{2}{e^{i\omega t}} \frac{L}{EI} \pi = \left\{ \int_0^L \left[ \sum_{i=1}^{N} B_i \frac{d\varphi_i(x)}{dx} \right] \ dx \right\ } + k_1 \left( \sum_{i=1}^{N} C_i \psi_i(x) \right)^2 + k_2 \left( \sum_{i=1}^{N} B_i \varphi_i(x) \right)^2 + k_3 \left( \sum_{i=1}^{N} C_i \psi_i(x) \right)^2 + k_4 \left( \sum_{i=1}^{N} B_i \varphi_i(x) \right)^2
\]

\[
- (\omega)^2 \left[ \int_0^L \left( \sum_{i=1}^{N} C_i \psi_i(x) \right)^2 \ dx + \int_0^L \left( \sum_{i=1}^{N} B_i \varphi_i(x) \right)^2 \ dx \right] \}
\]

(28)

The generalized displacement functions are adopted as polynomial functions of the form

\[
\psi_i(x) = (x)^{k-1} \Psi(x), \quad \varphi_i(x) = \frac{d\psi_i(x)}{dx}.
\]

(29)

for \( k=1 \) to \( k=N \), with

\[
\Psi(x) = \alpha_1 (x) + \alpha_2 (x)^2 + \alpha_3 (x)^3 + \alpha_4 (x)^4 + \alpha_5 (x)^5.
\]

(30)
where \( \alpha_1, \alpha_2, \alpha_3, \alpha_4 \) and \( \alpha_5 \), are constants.

Assuming \( \alpha_3 = 1 \), the others constants can be determined applying the boundary conditions at \( \bar{x} = 0 \) and \( \bar{x} = 1 \), respectively, as follows

\[
R \left. \frac{d^2 \Psi}{dx^2} \right|_{\bar{x}=0} = k_{i1} \left. \frac{d\Psi}{dx} \right|_{\bar{x}=0}, \quad R \left. \frac{d^2 \Psi}{dx^2} \right|_{\bar{x}=1} = -k_{i2} \left. \frac{d\Psi}{dx} \right|_{\bar{x}=1}
\]

(31)

\[2 \kappa \frac{\mu AL^2}{EI} \left. \frac{d\Psi}{dx} \right|_{\bar{x}=0} = -k_{i1} \Psi_0 \quad 2 \kappa \frac{\mu AL^2}{EI} \left. \frac{d\Psi}{dx} \right|_{\bar{x}=1} = k_{i2} \Psi_1 \]

or

\[
\frac{AL^2}{I} \left. \frac{\kappa}{1+\nu} \frac{d\Psi}{dx} \right|_{\bar{x}=0} = -k_{i1} \Psi_0 \quad \frac{AL^2}{I} \left. \frac{\kappa}{1+\nu} \frac{d\Psi}{dx} \right|_{\bar{x}=1} = k_{i2} \Psi_1
\]

(32)

Finally, replacing the approximate generalized displacements in Eq. (28), and minimizing this energy expression, a system of homogeneous equations of the first order for coefficients \( C_{ik} \) and \( B_{ik} \) is obtained:

\[
\frac{\partial \pi}{\partial c_k} = 0, \quad \frac{\partial \pi}{\partial b_k} = 0
\]

(33)

with \( k = 1, 2, \ldots, N \). The natural frequencies correspond to the solutions of the homogeneous form of linear algebraic equations, Eqs. (33), which can be written in matrix form as:

\[
[[\eta]] - \omega^2 [\lambda] = 0
\]

(34)

In order that at least one coefficient be different from zero, it is possible to equate the determinant of this system (34) to zero and obtain the frequency equation. The lowest solution which differs from zero represents the approximate frequency of the first mode, \( \omega_{-1} \), and the following solutions represent the higher frequencies, \( \omega_{-2}, \omega_{-3} \) and so on.

## 5 NUMERICAL RESULTS

In order to analyze the effect of non-classical supports on the free vibrations of Timoshenko micro beams the first three frequency coefficients are calculated for rectangular beam sections, \( A = b \times h, I = \frac{b \times h^3}{12}, \frac{AL^2}{I} = 12 \left( \frac{L}{h} \right)^2, \kappa = 5/6. \)

The Poisson’s ratio is taken \( \nu = 0.30 \) and \( R \) is assumed equal to 1 as in Ref. [2].

The number of terms is taken \( N=10 \) in all of the numerical examples, (the size of the determinant is 20).

Tables presented the first three frequency coefficients predicted by the present approach, for two different situations: a) neglecting the rotary inertia \( (m_2 = 0) \) and b) considering it \( (m_2 \neq 0) \).

The constants rigidities of the spring supports are taken: \( k_{i1} = k_{i2} \to \infty \), while \( k_{i1} = k_{i2} \) are varied from zero to infinity.

A comparison study is carried out between the present results of frequencies for Timoshenko micro beam with classical boundary conditions and the results given in the Ref. [2].

Tables 1, 2 and 3 show the frequency coefficients of free vibration of micro beams for three different length to thickness ratios \( L/h \): 20, 10 and 5.

The scale parameter ratio, \( l/h \), contains the internal material length scale parameter \( l \) and captures the size effect.
Table 1: First three natural frequency coefficients with and without rotary inertia \( \omega_j = \omega_j \sqrt{\frac{J A}{EI}} \), \( L = 20 \).

<table>
<thead>
<tr>
<th>( k_1 = k_2 \rightarrow \infty )</th>
<th>( \sqrt{\frac{l}{h}} )</th>
<th>( \omega - 1 )</th>
<th>( \omega - 2 )</th>
<th>( \omega - 3 )</th>
<th>( \text{Ref} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( k_1 = k_2 )</td>
<td>( m_2 = 0 )</td>
<td>( m_2 \neq 0 )</td>
<td>( m_2 = 0 )</td>
<td>( m_2 \neq 0 )</td>
<td>( m_2 = 0 )</td>
</tr>
<tr>
<td>0</td>
<td>9.838100</td>
<td>9.82813</td>
<td>38.98143</td>
<td>38.82991</td>
<td>86.36812</td>
</tr>
<tr>
<td></td>
<td>9.8381</td>
<td>9.8281</td>
<td>[2]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>11.63441</td>
<td>11.62265</td>
<td>40.90093</td>
<td>40.74256</td>
<td>88.27156</td>
</tr>
<tr>
<td>10</td>
<td>17.13222</td>
<td>17.11428</td>
<td>49.02354</td>
<td>48.85312</td>
<td>97.65056</td>
</tr>
<tr>
<td>10 (^3)</td>
<td>22.16475</td>
<td>22.13794</td>
<td>60.12550</td>
<td>59.86930</td>
<td>114.96777</td>
</tr>
<tr>
<td></td>
<td>22.02941</td>
<td>22.00293</td>
<td>[2]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \infty )</td>
<td>9.8381</td>
<td>9.8281</td>
<td>38.981</td>
<td>38.8299</td>
<td>86.36812</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>12.13486</td>
<td>12.12265</td>
<td>43.90093</td>
<td>43.74256</td>
</tr>
<tr>
<td></td>
<td>42.20885</td>
<td>42.17728</td>
<td>163.31817</td>
<td>162.91962</td>
<td>342.85031</td>
</tr>
<tr>
<td></td>
<td>42.20885</td>
<td>42.17728</td>
<td>[2]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>43.24897</td>
<td>43.21020</td>
<td>43.90093</td>
<td>43.74256</td>
<td>91.27156</td>
</tr>
<tr>
<td>10</td>
<td>46.18742</td>
<td>46.14753</td>
<td>63.90093</td>
<td>63.74256</td>
<td>134.96777</td>
</tr>
<tr>
<td>10 (^3)</td>
<td>81.70090</td>
<td>81.64484</td>
<td>212.92739</td>
<td>212.58392</td>
<td>389.89344</td>
</tr>
<tr>
<td></td>
<td>59.21265</td>
<td>59.16546</td>
<td>[2]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \infty )</td>
<td>43.24897</td>
<td>43.21020</td>
<td>43.90093</td>
<td>43.74256</td>
<td>91.27156</td>
</tr>
<tr>
<td>0</td>
<td>2.8384</td>
<td>58.92132</td>
<td>58.87605</td>
<td>217.04137</td>
<td>216.88389</td>
</tr>
<tr>
<td></td>
<td>58.92132</td>
<td>58.87605</td>
<td>[2]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>59.21265</td>
<td>59.16546</td>
<td>217.04137</td>
<td>216.88389</td>
<td>436.74807</td>
</tr>
<tr>
<td>10</td>
<td>61.19122</td>
<td>61.14425</td>
<td>217.04137</td>
<td>216.88389</td>
<td>436.74807</td>
</tr>
<tr>
<td>10 (^3)</td>
<td>106.95308</td>
<td>106.90467</td>
<td>217.04137</td>
<td>216.88389</td>
<td>436.74807</td>
</tr>
<tr>
<td></td>
<td>91.43261</td>
<td>91.36679</td>
<td>[2]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \infty )</td>
<td>80.57131</td>
<td>80.52044</td>
<td>279.80505</td>
<td>279.66241</td>
<td>531.07492</td>
</tr>
<tr>
<td>0</td>
<td>4</td>
<td>80.57131</td>
<td>80.52044</td>
<td>279.80505</td>
<td>279.66241</td>
</tr>
<tr>
<td></td>
<td>80.57131</td>
<td>80.52044</td>
<td>[2]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>80.74406</td>
<td>80.69322</td>
<td>279.80505</td>
<td>279.66241</td>
<td>531.07492</td>
</tr>
<tr>
<td>10</td>
<td>81.91179</td>
<td>81.86222</td>
<td>279.80505</td>
<td>279.66241</td>
<td>531.07492</td>
</tr>
<tr>
<td>10 (^3)</td>
<td>137.56396</td>
<td>137.53172</td>
<td>308.03344</td>
<td>307.95355</td>
<td>573.83252</td>
</tr>
<tr>
<td></td>
<td>153.38681</td>
<td>153.34874</td>
<td>[2]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \infty )</td>
<td>80.57131</td>
<td>80.52044</td>
<td>279.80505</td>
<td>279.66241</td>
<td>531.07492</td>
</tr>
<tr>
<td>0</td>
<td>4.7221</td>
<td>88.80009</td>
<td>88.75037</td>
<td>300.76328</td>
<td>300.68412</td>
</tr>
<tr>
<td></td>
<td>88.80009</td>
<td>88.75037</td>
<td>[2]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>88.91848</td>
<td>88.89226</td>
<td>300.76328</td>
<td>300.68412</td>
<td>558.94459</td>
</tr>
<tr>
<td>10</td>
<td>89.86225</td>
<td>89.83785</td>
<td>300.76328</td>
<td>300.68412</td>
<td>558.94459</td>
</tr>
<tr>
<td>10 (^3)</td>
<td>148.10283</td>
<td>148.08443</td>
<td>354.67003</td>
<td>354.65920</td>
<td>600.17563</td>
</tr>
<tr>
<td></td>
<td>164.06086</td>
<td>164.04626</td>
<td>[2]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2: First three natural frequency coefficients with and without rotary inertia $\omega_j = \omega_j L^2 \sqrt{\frac{\rho A}{EI}}$, $\sqrt{\frac{L}{h}} = 10$

<table>
<thead>
<tr>
<th>$k_{1} = k_{2}$</th>
<th>$\sqrt[3]{h}$</th>
<th>$\omega_{-1}$</th>
<th>$\omega_{-2}$</th>
<th>$\omega_{-3}$</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_{1} = k_{2}$</td>
<td>$m_2 = 0$</td>
<td>$m_2 \neq 0$</td>
<td>$m_2 = 0$</td>
<td>$m_2 \neq 0$</td>
<td>$m_2 = 0$</td>
</tr>
<tr>
<td>0 0</td>
<td>9.74536</td>
<td>9.70748</td>
<td>37.59640</td>
<td>37.09616</td>
<td>80.06125</td>
</tr>
<tr>
<td></td>
<td>9.7454</td>
<td>9.7075</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>11.36266</td>
<td>11.31921</td>
<td>39.17823</td>
<td>38.66577</td>
<td>81.43688</td>
</tr>
<tr>
<td>10</td>
<td>16.63407</td>
<td>16.57202</td>
<td>46.32031</td>
<td>45.75058</td>
<td>88.42982</td>
</tr>
<tr>
<td>$10^3$</td>
<td>21.18711</td>
<td>21.10187</td>
<td>55.53070</td>
<td>54.84877</td>
<td>99.76608</td>
</tr>
<tr>
<td>$\infty$</td>
<td>21.08441</td>
<td>20.99943</td>
<td>55.40106</td>
<td>54.71943</td>
<td>99.75113</td>
</tr>
<tr>
<td>0 1</td>
<td>22.82705</td>
<td>22.74780</td>
<td>85.54710</td>
<td>84.81459</td>
<td>176.04627</td>
</tr>
<tr>
<td></td>
<td>22.8270</td>
<td>22.7478</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>23.52233</td>
<td>23.44238</td>
<td>86.01666</td>
<td>85.29107</td>
<td>176.28559</td>
</tr>
<tr>
<td>10</td>
<td>27.45430</td>
<td>27.37111</td>
<td>89.00940</td>
<td>88.33459</td>
<td>177.77706</td>
</tr>
<tr>
<td>$10^3$</td>
<td>43.30640</td>
<td>43.20386</td>
<td>110.82005</td>
<td>110.25961</td>
<td>198.18585</td>
</tr>
<tr>
<td>$\infty$</td>
<td>47.81062</td>
<td>47.69064</td>
<td>122.24100</td>
<td>121.63312</td>
<td>214.12094</td>
</tr>
<tr>
<td>0 1.41421</td>
<td>30.39601</td>
<td>30.30227</td>
<td>110.35968</td>
<td>109.76364</td>
<td>219.42483</td>
</tr>
<tr>
<td></td>
<td>30.3960</td>
<td>30.3023</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>30.86674</td>
<td>30.77329</td>
<td>110.58893</td>
<td>110.00054</td>
<td>219.48507</td>
</tr>
<tr>
<td>10</td>
<td>33.74870</td>
<td>33.65808</td>
<td>112.10021</td>
<td>111.56178</td>
<td>219.84400</td>
</tr>
<tr>
<td>$10^3$</td>
<td>54.71426</td>
<td>54.62751</td>
<td>137.09715</td>
<td>136.78827</td>
<td>242.10232</td>
</tr>
<tr>
<td>$\infty$</td>
<td>61.31955</td>
<td>61.22567</td>
<td>152.37133</td>
<td>152.10957</td>
<td>258.99148</td>
</tr>
<tr>
<td>0 1.73205</td>
<td>36.12971</td>
<td>36.03056</td>
<td>127.47679</td>
<td>127.05876</td>
<td>246.57493</td>
</tr>
<tr>
<td></td>
<td>36.1297</td>
<td>36.0305</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>36.48387</td>
<td>36.38552</td>
<td>127.59831</td>
<td>127.18565</td>
<td>246.58676</td>
</tr>
<tr>
<td>10</td>
<td>38.72450</td>
<td>38.63192</td>
<td>128.40185</td>
<td>128.02462</td>
<td>246.63184</td>
</tr>
<tr>
<td>$10^3$</td>
<td>62.86248</td>
<td>62.79548</td>
<td>154.73071</td>
<td>154.59598</td>
<td>269.64548</td>
</tr>
<tr>
<td>$\infty$</td>
<td>70.47708</td>
<td>70.41245</td>
<td>171.20632</td>
<td>171.12363</td>
<td>285.05341</td>
</tr>
<tr>
<td>0 2</td>
<td>40.82954</td>
<td>40.72991</td>
<td>140.39116</td>
<td>140.12740</td>
<td>265.54739</td>
</tr>
<tr>
<td></td>
<td>40.8295</td>
<td>40.7299</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>41.10902</td>
<td>41.01044</td>
<td>140.45580</td>
<td>140.19530</td>
<td>265.55268</td>
</tr>
<tr>
<td>10</td>
<td>42.90907</td>
<td>42.81767</td>
<td>140.87643</td>
<td>140.63762</td>
<td>265.55619</td>
</tr>
<tr>
<td>$10^3$</td>
<td>69.18760</td>
<td>69.13880</td>
<td>167.73418</td>
<td>167.69199</td>
<td>288.88632</td>
</tr>
<tr>
<td>$\infty$</td>
<td>77.28501</td>
<td>77.24353</td>
<td>184.35011</td>
<td>184.30782</td>
<td>302.43612</td>
</tr>
<tr>
<td>0 2.2336</td>
<td>44.83858</td>
<td>44.74138</td>
<td>150.61720</td>
<td>150.46969</td>
<td>279.66316</td>
</tr>
<tr>
<td></td>
<td>44.8386</td>
<td>44.7414</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>45.06514</td>
<td>44.96911</td>
<td>150.65028</td>
<td>150.50451</td>
<td>279.67807</td>
</tr>
<tr>
<td>10</td>
<td>46.54061</td>
<td>46.45234</td>
<td>150.85561</td>
<td>150.72142</td>
<td>279.74716</td>
</tr>
<tr>
<td>$10^3$</td>
<td>74.31528</td>
<td>74.28150</td>
<td>177.83100</td>
<td>177.81743</td>
<td>303.18744</td>
</tr>
<tr>
<td>$\infty$</td>
<td>82.60804</td>
<td>82.58246</td>
<td>194.11597</td>
<td>194.02462</td>
<td>314.98102</td>
</tr>
<tr>
<td>$k_1 = k_2 \rightarrow \infty$</td>
<td>$\frac{l}{h}$</td>
<td>$\omega_1$</td>
<td>$\omega_2$</td>
<td>$\omega_3$</td>
<td>$\omega_4$</td>
</tr>
<tr>
<td>----------------</td>
<td>--------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>$k_1 = k_2$</td>
<td>$m_2 = 0$</td>
<td>$m_2 \neq 0$</td>
<td>$m_2 = 0$</td>
<td>$m_2 \neq 0$</td>
<td>$m_2 = 0$</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>9.39901</td>
<td>9.27404</td>
<td>33.24004</td>
<td>32.16650</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9.3990</td>
<td>9.2740</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>10.81705</td>
<td>10.68212</td>
<td>34.31247</td>
<td>33.25863</td>
<td>64.71882</td>
</tr>
<tr>
<td>10</td>
<td>15.07239</td>
<td>14.92140</td>
<td>38.83469</td>
<td>37.88338</td>
<td>67.62210</td>
</tr>
<tr>
<td>$10^3$</td>
<td>18.25187</td>
<td>18.09058</td>
<td>43.94473</td>
<td>43.08844</td>
<td>71.15855</td>
</tr>
<tr>
<td>$10^6$</td>
<td>18.23206</td>
<td>18.06878</td>
<td>44.00308</td>
<td>43.13137</td>
<td>70.22770</td>
</tr>
<tr>
<td>$\infty$</td>
<td>18.23425</td>
<td>18.07066</td>
<td>31.20904</td>
<td>31.21811</td>
<td>43.53406</td>
</tr>
<tr>
<td>0</td>
<td>0.5000</td>
<td>13.70567</td>
<td>13.54211</td>
<td>48.12664</td>
<td>47.05523</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13.7057</td>
<td>13.5421</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>14.63809</td>
<td>14.47451</td>
<td>48.63274</td>
<td>47.60239</td>
<td>93.57124</td>
</tr>
<tr>
<td>10</td>
<td>18.28605</td>
<td>18.13670</td>
<td>51.17697</td>
<td>50.34881</td>
<td>94.49277</td>
</tr>
<tr>
<td>$10^3$</td>
<td>24.66317</td>
<td>24.51885</td>
<td>59.86368</td>
<td>59.27227</td>
<td>102.48074</td>
</tr>
<tr>
<td>$10^6$</td>
<td>26.71225</td>
<td>26.54505</td>
<td>65.08995</td>
<td>64.44208</td>
<td>110.11762</td>
</tr>
<tr>
<td>$\infty$</td>
<td>26.71337</td>
<td>26.54612</td>
<td>65.09351</td>
<td>64.46622</td>
<td>110.12336</td>
</tr>
<tr>
<td>0</td>
<td>0.7071</td>
<td>16.80112</td>
<td>16.62134</td>
<td>57.90759</td>
<td>57.04939</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16.8011</td>
<td>16.6213</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>17.49450</td>
<td>17.31908</td>
<td>58.18157</td>
<td>57.35936</td>
<td>111.11550</td>
</tr>
<tr>
<td>10</td>
<td>20.47203</td>
<td>20.32420</td>
<td>59.62936</td>
<td>58.98491</td>
<td>111.37250</td>
</tr>
<tr>
<td>$10^3$</td>
<td>29.02810</td>
<td>28.90847</td>
<td>70.17078</td>
<td>69.82572</td>
<td>121.73340</td>
</tr>
<tr>
<td>$10^6$</td>
<td>32.11873</td>
<td>31.98829</td>
<td>77.53265</td>
<td>77.23382</td>
<td>130.93515</td>
</tr>
<tr>
<td>$\infty$</td>
<td>32.12081</td>
<td>31.99036</td>
<td>77.53733</td>
<td>77.23857</td>
<td>130.93783</td>
</tr>
<tr>
<td>0</td>
<td>0.8660</td>
<td>19.28774</td>
<td>19.10288</td>
<td>65.20190</td>
<td>64.58787</td>
</tr>
<tr>
<td></td>
<td></td>
<td>19.2877</td>
<td>19.1029</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>19.83217</td>
<td>19.65388</td>
<td>65.3539</td>
<td>64.76577</td>
<td>123.46977</td>
</tr>
<tr>
<td>10</td>
<td>22.92919</td>
<td>22.14856</td>
<td>66.17810</td>
<td>65.71894</td>
<td>123.50511</td>
</tr>
<tr>
<td>$10^3$</td>
<td>32.42533</td>
<td>32.33172</td>
<td>77.88703</td>
<td>77.72564</td>
<td>135.42586</td>
</tr>
<tr>
<td>$10^6$</td>
<td>36.11020</td>
<td>36.01854</td>
<td>86.26111</td>
<td>86.17076</td>
<td>144.93020</td>
</tr>
<tr>
<td>$\infty$</td>
<td>36.11252</td>
<td>36.02087</td>
<td>86.26534</td>
<td>86.17505</td>
<td>144.92881</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>21.38700</td>
<td>21.20365</td>
<td>70.96763</td>
<td>70.57004</td>
</tr>
<tr>
<td>1</td>
<td>21.82742</td>
<td>21.65168</td>
<td>71.04977</td>
<td>70.66859</td>
<td>132.78899</td>
</tr>
<tr>
<td>10</td>
<td>23.88417</td>
<td>23.74634</td>
<td>71.50218</td>
<td>71.20343</td>
<td>132.79196</td>
</tr>
<tr>
<td>$10^3$</td>
<td>35.20362</td>
<td>35.13392</td>
<td>83.99424</td>
<td>83.94271</td>
<td>145.83037</td>
</tr>
<tr>
<td>$10^6$</td>
<td>39.24110</td>
<td>39.18156</td>
<td>92.83294</td>
<td>92.80403</td>
<td>155.15966</td>
</tr>
<tr>
<td>$\infty$</td>
<td>39.24332</td>
<td>39.18381</td>
<td>92.83617</td>
<td>92.80718</td>
<td>155.15433</td>
</tr>
<tr>
<td>0</td>
<td>1.1180</td>
<td>23.21094</td>
<td>23.03320</td>
<td>75.68792</td>
<td>75.45988</td>
</tr>
<tr>
<td></td>
<td></td>
<td>23.2109</td>
<td>23.0332</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>23.57384</td>
<td>23.40410</td>
<td>75.72917</td>
<td>75.51021</td>
<td>140.12420</td>
</tr>
<tr>
<td>10</td>
<td>25.30897</td>
<td>25.17805</td>
<td>75.95814</td>
<td>75.78525</td>
<td>140.18552</td>
</tr>
<tr>
<td>$10^3$</td>
<td>35.54310</td>
<td>37.49374</td>
<td>88.99309</td>
<td>88.98323</td>
<td>154.06620</td>
</tr>
<tr>
<td>$10^6$</td>
<td>41.78783</td>
<td>41.75190</td>
<td>97.99937</td>
<td>97.91748</td>
<td>163.02548</td>
</tr>
<tr>
<td>$\infty$</td>
<td>41.78984</td>
<td>41.75392</td>
<td>98.00141</td>
<td>97.91924</td>
<td>163.01660</td>
</tr>
</tbody>
</table>

The results obtained from the present study are in excellent agreement with those of Araújo dos Santos and Reddy [2] for classical boundary conditions, as observed in Tables 1, 2 and 3.
2 CONCLUSIONS
- It can be observed that the effect of the rotatory inertia decreases the frequency coefficients of free vibration of the considered micro beams, as well as that its influence decreases as $l/h$ increases.
- For large length to thickness ratios, the rotary inertia has a negligible effect.
- It is evident from the obtained values of frequency parameters that, when the rigidities of the springs are taken as $k_1 = k_3 = k_2 \to \infty$, then, the beam can be considered as a beam clamped at the both ends. When they are taken as $k_1 = k_2 \to \infty$ and $k_1 = k_2 = 0$ the beam can be considered a simply supported beam.
- Increment in the values of rotational rigidities, $k_1 = k_2$, is more noticeable on the fundamental frequency coefficients of the micro beam than the second and third frequency coefficients.
- It is observed from the investigations that, all of the obtained results are very accurate and may be useful to other researchers so as to compare their results.

AKNOWLEDGEMENTS
The authors wish to express their gratitude to the Universidad Nacional del Sur, the Consejo Nacional de Investigaciones Científicas y Técnicas and Comisión de Investigaciones Científicas de la Provincia de Buenos Aires for the financial support, which enable this work to be done.

REFERENCES


ANALYSIS OF THE BEHAVIOR OF AN INDUSTRIAL BUILDING WITH PRESENCE OF VIBRATIONS DUE TO MACHINERY

ERCOLI N.L.¹*, PERALTA M. H.², PICO L.³

¹ Universidad Nacional del Centro de la Provincia de Buenos Aires, Facultad de Ingeniería Avda. Del Valle 5737. 7400 Olavarría. Bs. As. Argentina E-mail: nercoli@fio.unicen.edu.ar. http://www.fio.unicen.edu.ar

² Universidad Nacional del Centro de la Provincia de Buenos Aires, Facultad de Ingeniería Avda. Del Valle 5737. 7400 Olavarría. Bs.As. Argentina E-mail: mperalta@fio.unicen.edu.ar. http://www.fio.unicen.edu.ar

³ Universidad Nacional del Centro de la Provincia de Buenos Aires, Facultad de Ingeniería Avda. Del Valle 5737. 7400 Olavarría. Bs.As. Argentina E-mail: lpico@fio.unicen.edu.ar. http://www.fio.unicen.edu.ar

Keywords: dynamic analysis, FEM, machine vibrations

Abstract

The design of structures in industrial buildings is complex due to the diversity and nature of the processes involved according to the destination with presence of systems and equipment that generate vibrations. In order to achieve a suitable technical solution, the designer must possess adequate knowledge of every aspect related both to the equipment or machine and to the characteristics of the structure the equipment will be linked to so as to mitigate vibration problems.

This paper addresses the issues related to a building which supports a crusher and screen equipment in a cement plant facilities in the city of Olavarría, Province of Buenos Aires. The building is a reinforced concrete structure composed of slabs and spatial frames laying on a concrete slab foundation.

Computer simulation models were calibrated and used to determine and analyze the structure dynamic properties and the structural response under service load. Design criteria are established in order to achieve a proper structural behavior, as well as recommendations related to the use of vibration isolation, obtained from the analysis results.
1. INTRODUCTION

The structural design of industrial buildings is a complex problem associated with the range and nature of the processes involved according to their destination with the presence of equipment requiring an acceptable structural response. In our environment, where important industries have been established, the design of industrial buildings requires the analysis and decision making concerning their appropriate behavior in the presence of equipment and systems that generate vibrations [1], [2]. The mechanical vibrations caused by industrial equipment affect the structural behavior. All aspects can be condensed into one objective: the modification of the global response to improve the structural performance. [3], [4]. The two ways to achieve appropriate behavior would be, on the one hand, to detune the structure affected by vibrations and, on the other hand, to act on the machine-structure link through the insulation system design modifying the transmissibility [5].

The structural design that involves the presence of mechanical systems requires an overall view on the machine-structure interaction, which in some cases can extend to the foundation soil, [6], [7] and [8]. The interpretation of the origin of mechanical vibrations and their effects on structural components is a subject of analysis and, in general, it requires experimental validation measurements.

In any vibration problem three essential components can be identified:
• Machinery in operation and vibration originating source.
• Supporting structure of the machinery (floor, foundation, bank, etc.).
• Vibration assembly interposed between both systems, the structure and the machine.

Considering the above, the forms of assistance to mitigate vibration problems depend on the operating conditions and design limitations. The options are:
• From a purely mechanical point of view, the intervention on the source that produces vibration (machine or mechanism) trying to minimize the generation of vibratory forces, for example through rotor balancing.
• From a purely structural point of view, intervention on the structure affected by vibration. The options with the aim of detuning are: modification of mass or stiffness to change their natural frequencies, and / or increased damping to minimize the amplitude of the oscillations, [7] and [8].
• From a structural-mechanical point of view, intervention on the link between the machine and its foundation by vibration isolation techniques, i.e., by modifying the transmissibility between the two systems [3].

In particular, for a technically suitable solution, the designer is required to have adequate knowledge of all aspects related with the equipment or machine, and the characteristics of the structure the machine will be connected to.

In this paper, an industrial building located in a cement plant for the process of classification and crushing of limestone through an industrial sieve and a crusher with a hopper is analyzed. The building design, plan and height are the result of meeting the requirements related to the aforesaid mechanical systems and their operation. A dynamic analysis of the structure is performed in order to produce its dynamic characteristics in terms of modes and frequencies.
On the other hand, the reinforced concrete floor responses where unloading machines work with a corresponding dynamic load are analyzed. This information will permit the assessment of the structural behavior under service conditions and decision making based on the results [3] [4].

2 THEORETICAL ASPECTS AND GUIDELINES OF STRUCTURAL PERFORMANCE EVALUATION

The dynamic characteristics, modes and frequencies of vibration, can be taken as indicators of structural behavior, including the possible involvement of bearing capacity, given the parameters involved in them, mass, M, stiffness, and damping K, C [9], according to equation 1, which governs the problems of structural dynamics,

\[ M \ddot{A} + C \dot{V} + K D = F(t) \]  \hspace{1cm} (1)

Figure 1 shows the important role of damping and the ratio between the excitation frequency and the fundamental frequency on the response by varying the dynamic amplification factor.

The structural response obtained through experimental and/or numerical measurements in terms of amplitudes of accelerations and velocities, allows the performance evaluation by comparing the response to that indicated by the standards and criteria taken as reference [3] [4].

To establish an evaluation criterion for structural damage there is a set of standards and criteria which express the allowable values and vibratory maximum limits that could damage a structure. For the purposes of this analysis, the criteria established in international standards (ISO, DIN) shown in Figure 2, according to [3] are adopted. It is observed that the values expressed in units of acceleration, m/s², as well as levels of vibration velocity in mm/s, rms, are related to the vibrational frequency and the damage that these levels may cause on structures.
Other authors, such as Rades [4], show other criteria according to Table 1, with values consistent with Figure 2.

**Table 1: Vibration Criteria**

<table>
<thead>
<tr>
<th>Range</th>
<th>Velocity rms (mm/s)</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Less than 2.5</td>
<td>no damage occurs</td>
</tr>
<tr>
<td>II</td>
<td>2.5 to 5.0</td>
<td>highly improbable damage</td>
</tr>
<tr>
<td>III</td>
<td>5.0 to 10.0</td>
<td>unlikely damage</td>
</tr>
<tr>
<td>IV</td>
<td>Greater than 10.0</td>
<td>possible damage; necessary structural assessment</td>
</tr>
</tbody>
</table>

As the chart shows in Figure 2 and Table 1, the acceptable value is taken as not generating structural damage when 10 mm/s, rms, of vibration velocity, and the acceleration level of 0.1 m/s² are not exceeded.

### 3. APPLICATION: INDUSTRIAL BUILDING WITH EQUIPMENTS OF SCREEN AND CRUSHER INSTALLED

#### 3.1. Information available

- Memory of calculation of the structure and its foundation
- Report with characteristics of soil foundation
- Report with the characteristics of the equipment to be installed
- Report of Predictive Maintenance with measurements relative to the vibratory equipment installed in the building of similar characteristics in service.
3.2 Description of the structural typology
It is a reinforced concrete structure built on site; a multi-level spatially framed structure of plan dimensions 11.20 m x 7.30 m with columns at each corner. It has a reinforced concrete deck area where conveyor belt supports of limestone supply are discharged. There are three levels with reinforced concrete slabs supported on beams, with large sections of openings due to the requirements of the equipment to be installed. The building is founded by a plate of reinforced concrete, of rigid structural behavior of 0.60 m thickness (reference level 195.49 m). The total height of the building is 19.90 m. The + 203.85 m level corresponds to the crusher hopper support, and the level + 206.85 to the screen; details are shown in Figures 3 and 4. The design of the dimensions of the different structural steel and concrete elements were made according to CIRSOC 201/05 [10]. The qualities of the materials used are H-30 concrete and 420 DNA steel.

3.3 Characteristics of equipment and machines

Screen
The screen at 206.850 m level, Figure 5, is supported on the reinforced concrete slab at four points, with shares contributed by the manufacturer /supplier of the equipment. At each point the value is: static load of 10.42 kN and dynamic load of +1.372 kN. The characteristic data of the engine are power 30 kW, velocity 996 rpm (frequency 15 Hz) and voltage 380/660 V.
Crusher and Hopper
Crusher and hopper are at 203.850 m level. Figure 6 shows a detail of the crusher and discharge points in the structure of the reinforced concrete slab floor with values according to Table 2. The characteristics of the motor 960 rpm, or 15 HZ

Table 2: Load values in crusher support

<table>
<thead>
<tr>
<th>PE500*750</th>
<th>Static Load(kN)</th>
<th>Dynamic load(kN)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>18.4</td>
<td>±7</td>
</tr>
<tr>
<td>B</td>
<td>18.4</td>
<td>±5.5</td>
</tr>
<tr>
<td>C</td>
<td>18.4</td>
<td>±4</td>
</tr>
</tbody>
</table>

The hopper, as can be seen in Figure 7, supported by 4 points on the concrete slab. The full hopper condition weighs 110 t.
3.4 NUMERICAL ANALYSIS

3.4.1 Evaluation of dynamic characteristics of the building
The overall analysis in order to analyze the impact on the structural behavior of the installation of equipment and crusher screen is carried out.

3.4.1.a Models analyzed
For later analysis by the Finite Element Method, models of the reinforced concrete building dimensions were raised from drawings that include founding by an audience of concrete of 0.60 m thick and considering foundation embedment on improved soil with the addition of 4 m thick of clinker (Figure 8). The brick element available in the library of ALGOR software was used [11]

3.4.1.b Dynamic Analysis
An analysis of the dynamic characteristics (vibration modes and frequencies) of the building structure was performed, considering the first 15 modes of vibration for the following service
load conditions:
Case a. Performance under building dead load.
Case b. Performance under building dead load, and the masses of the machines according to
reports delivered by the company.
Case c. Same condition as Case b, adding the mass under full hopper condition in level
+203.850.

3.4.1.c Results Analysis
Table 2 indicates the results in terms of natural frequencies for each of the building 15 modes
of vibration considering the three cases previously mentioned. In addition, mode type for each
case is indicated.

Table 2: Natural Frequencies $f$, in Hz.

<table>
<thead>
<tr>
<th>Mode</th>
<th>Structure without machine, Case a</th>
<th>Structure with machine, Case b</th>
<th>Structure with machine, and full hopper, Case c</th>
<th>Mode Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.11</td>
<td>2.11</td>
<td>2.11</td>
<td>Flexural cantilever $yz$</td>
</tr>
<tr>
<td>2</td>
<td>3.08</td>
<td>3.08</td>
<td>3.08</td>
<td>Flexural cantilever $xz$</td>
</tr>
<tr>
<td>3</td>
<td>3.77</td>
<td>3.77</td>
<td>3.77</td>
<td>Torsional according to $z$</td>
</tr>
<tr>
<td>4</td>
<td>5.80</td>
<td>5.80</td>
<td>5.80</td>
<td>Flexural $yz$</td>
</tr>
<tr>
<td>5</td>
<td>8.94</td>
<td>8.94</td>
<td>8.93</td>
<td>Flexural $xz$</td>
</tr>
<tr>
<td>6</td>
<td>10.62</td>
<td>10.62</td>
<td>10.61</td>
<td>Flexural beams, level 215.300 y 211.150</td>
</tr>
<tr>
<td>7</td>
<td>11.20</td>
<td>11.20</td>
<td>11.20</td>
<td>Torsional according to $z$</td>
</tr>
<tr>
<td>8</td>
<td>15.04</td>
<td>13.97</td>
<td>13.14</td>
<td>Flexural beams $xy$ level 211.150</td>
</tr>
<tr>
<td>9</td>
<td>15.77</td>
<td>14.81</td>
<td>14.64</td>
<td>Flexural beams $xy$ level 211.150</td>
</tr>
<tr>
<td>10</td>
<td>16.00</td>
<td>15.62</td>
<td>14.81</td>
<td>Flexural $xz$ beams, slab and columns</td>
</tr>
<tr>
<td>11</td>
<td>18.38</td>
<td>15.64</td>
<td>15.63</td>
<td>Flexural $xz$ beams, slab and columns</td>
</tr>
<tr>
<td>12</td>
<td>19.44</td>
<td>16.62</td>
<td>15.67</td>
<td>Flexural $z$, beams, slab and columns</td>
</tr>
<tr>
<td>13</td>
<td>19.64</td>
<td>19.38</td>
<td>16.61</td>
<td>Flexural slab level 206.850</td>
</tr>
<tr>
<td>14</td>
<td>22.06</td>
<td>19.50</td>
<td>17.22</td>
<td>Axial and flexural slab, level 203.850</td>
</tr>
<tr>
<td>15</td>
<td>23.32</td>
<td>23.22</td>
<td>19.42</td>
<td>Global Structure</td>
</tr>
</tbody>
</table>

Figure 9 shows the vibration mode 1 for conditions Case a, b and c.

Figure 9: Vibration Mode 1
3.4.2 Evaluation of the dynamic characteristics of the floors where the machines will be installed.

3.4.2.a Analyzed Models
Models were proposed, for later analysis using Finite Element Method (FEM), corresponding to reinforced concrete floor (beams and slab) levels +203.850 m and 206.850 m, where the crusher and screen machines will be installed. Dimensions were considered as shown in general arrangement plans, including the corresponding geometric and discontinuities.

3.4.2.b Dynamic Analysis: Modes and Frequencies
Analysis of the dynamic characteristics (vibrational modes and frequencies) of each reinforced concrete floor was performed. The first ten vibration modes were considered:
Case a. Reinforced concrete floor (slab and beams), without machines (for both levels)
Case b. Reinforced concrete floor (slab and beams), with machines (for both levels)
Case c. Reinforced concrete floor (slab and beams), with machines and full hopper (level +203.850m).

3.4.2.c Analysis Results
Tables 3 and 4 show the results corresponding to the ten vibration modes for each of the three conditions analyzed. Mode type for each case is also indicated.

Table 3: Natural Frequencies in H.(f)

<table>
<thead>
<tr>
<th>LEVEL 203,850 m – CRUSHED and HOPPER</th>
<th>Mode</th>
<th>f(Hz) Case a</th>
<th>f(Hz)Case b</th>
<th>f(Hz)Case c</th>
<th>f Case a /f Case b</th>
<th>f Case a/f Case c</th>
<th>Mode Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22.60</td>
<td>18.56</td>
<td>12.06</td>
<td>1.22</td>
<td>1.87</td>
<td>Flexural vertical</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>25.71</td>
<td>25.61</td>
<td>15.00</td>
<td>1.00</td>
<td>1.71</td>
<td>Flexural horizontal</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>29.49</td>
<td>26.38</td>
<td>22.53</td>
<td>1.12</td>
<td>1.31</td>
<td>Flexural vertical cantilever</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>34.27</td>
<td>29.34</td>
<td>27.20</td>
<td>1.17</td>
<td>1.26</td>
<td>Flexural vertical cantilever</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>35.71</td>
<td>35.21</td>
<td>27.82</td>
<td>1.01</td>
<td>1.28</td>
<td>Flexural vertical</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>39.43</td>
<td>36.83</td>
<td>30.47</td>
<td>1.07</td>
<td>1.29</td>
<td>Flexural vertical cantilever</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>43.03</td>
<td>39.36</td>
<td>33.47</td>
<td>1.09</td>
<td>1.29</td>
<td>Flexural vertical cantilever</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>43.22</td>
<td>42.90</td>
<td>39.33</td>
<td>1.01</td>
<td>1.10</td>
<td>Flexural vertical cantilever</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>46.35</td>
<td>46.22</td>
<td>43.06</td>
<td>1.00</td>
<td>1.08</td>
<td>Flexural vertical</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>47.97</td>
<td>46.83</td>
<td>43.29</td>
<td>1.02</td>
<td>1.11</td>
<td>Flexural horizontal</td>
<td></td>
</tr>
</tbody>
</table>

Ercoli, N; Peralta, M.L; Pico,L
Table 4: Natural Frequencies in Hz(f)

<table>
<thead>
<tr>
<th>MODE</th>
<th>f(Hz) Case a</th>
<th>f(Hz) Case b</th>
<th>f Case a/f Case b</th>
<th>Type Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16.03</td>
<td>12.16</td>
<td>1.32</td>
<td>flexural vertical</td>
</tr>
<tr>
<td>2</td>
<td>17.35</td>
<td>13.28</td>
<td>1.31</td>
<td>flexural vertical</td>
</tr>
<tr>
<td>3</td>
<td>26.25</td>
<td>25.10</td>
<td>1.05</td>
<td>flexural vertical</td>
</tr>
<tr>
<td>4</td>
<td>28.40</td>
<td>25.56</td>
<td>1.11</td>
<td>flexural horizontal</td>
</tr>
<tr>
<td>5</td>
<td>33.51</td>
<td>28.05</td>
<td>1.19</td>
<td>flexural vertical</td>
</tr>
<tr>
<td>6</td>
<td>35.98</td>
<td>28.25</td>
<td>1.27</td>
<td>flexural vertical</td>
</tr>
<tr>
<td>7</td>
<td>37.83</td>
<td>34.75</td>
<td>1.09</td>
<td>flexural vertical</td>
</tr>
<tr>
<td>8</td>
<td>38.27</td>
<td>37.25</td>
<td>1.03</td>
<td>flexural vertical</td>
</tr>
<tr>
<td>9</td>
<td>44.63</td>
<td>43.06</td>
<td>1.04</td>
<td>flexural horizontal</td>
</tr>
<tr>
<td>10</td>
<td>47.14</td>
<td>43.13</td>
<td>1.09</td>
<td>Torsional</td>
</tr>
</tbody>
</table>

Figures 10 and 11 show the first vibration mode, for each reinforced concrete floor.

![Vibration Mode 1 of reinforced concrete floor level +203.850 m](image1)

**Figure 10:** Vibration Mode 1 of reinforced concrete floor level +203.850 m

![Vibration Mode 1 of reinforced concrete floor +206.850 m](image2)

**Figure 11:** Vibration Mode 1 of reinforced concrete floor +206.850 m

3.4.3 Structural Response Evaluation of concrete floor levels +203.850m and +206.850 m

3.4.3.1 Models Analyzed
The same geometric models for analysis as those used in the previous section were considered. The corresponding action was applied on the machinery location, according to available data of the dynamic load depending on the frequency.

3.4.3.b Analysis
A corresponding analysis was performed while evaluating strains and stresses.

3.4.3.c Results
Figures 12 and 13 show displacements and principal maximum tensile stresses, in N/mm², in levels + 203.850 and +206.85, respectively.

4. RESULTS ANALYSIS
The global structure presents some vibration modes with frequencies near the natural frequency of excitation that machines contribute to. In the particular case of the floors analyzed (levels +203.850 and +206.850), the first modes have frequencies close to the frequencies of excitation (crusher and screen) values. This situation suggests that vibrational amplitudes may occur above the acceptable values in international standards.

In evaluating the response of the slabs having significant geometric discontinuities (holes) and areas of high flexibility for abrupt changes in cross sections, values indicating dynamic behaviors close to acceptable limits are observed.

5. Conclusions
At this stage of analysis, from the evaluation of computational numerical models, analytical predictions have been made of the dynamic characteristics (modes and frequencies) of the
global structure and the floor reinforced concrete structure, levels +203.850 and +206.850, where the crusher and screen will be implemented respectively. In the case of the floors, the prediction of structural behavior has also been made in service condition of machines. This information allows for the evaluation of the structural behavior of the global structure and the floors, corresponding to level Crusher and Screen. The results allow us to infer that displacement amplification phenomena and unsuitable speeds may occur.

In the building actual behavior with machines operating simultaneously, coupling phenomena and/or influence of mounting conditions may also occur. As these are not included in the analytical model, an evaluation is suggested from the dynamic measurements on site in the points of interest in the structure stemming from the analytical analysis.

Based on the results of analytical studies and on site measurements, it will be possible to analyze the structural behavior conditions and, considering the limits set by international standards, the study to define the placement of anti-vibrational mounts could be completed to achieve a proper structural performance.

REFERENCES
[10] INTI-CIRSOC, CIRSOC 201, 2005- Reglamento Argentino de Estructuras de Hormigón

Acknowledgements
The authors would like to thank the Engineering Company Cementos Avellaneda SA, for facilitating the use of data and information contained in this work.
A SIMPLIFIED NUMERICAL MODEL FOR ASSESING THE RESIDUAL STRENGTH AND DEFORMATION CAPACITY OF ASYMMETRICALLY DAMAGED ROPES

JUAN F. BELTRAN¹ AND ENZO DE VICO²

¹ Civil Engineering Department – University of Chile
jbeltran@ing.uchile.cl

² Civil Engineering Department – University of Chile
enzodevico@gmail.com

Key words: Damaged Ropes, Asymmetric Damage, Analytical Model.

Abstract. A simplified nonlinear mechanical model is proposed to estimate the residual strength and deformation capacity of ropes with asymmetric damage distribution. In this study, damage corresponds to the complete rupture of one or more rope components that belong to the outermost layer of a particular rope cross-section. In the proposed model, the damaged rope is assumed to behave as a nonlinear beam under biaxial bending and axial load with Bernoulli’s kinematic hypothesis. Biaxial bending arises from the unbalanced radial contact forces within rope cross-section, which are related to the initial helical geometry configuration of the rope components, due to the asymmetric damage distribution. An iterative cross-sectional numerical algorithm is implemented to estimate the asymmetric damaged rope capacity curve, stress and strain distributions throughout rope cross-section and rope geometry deformation for a prescribed axial displacement of the rope. This model is capable of predicting the strain gradient on a given asymmetric damaged rope cross-section in which the strains concentrate in rope components adjacent to the broken ones. This strain concentration can cause the premature failure of rope components and reduce the maximum load and failure axial strain that a damaged rope is capable of resisting relative to that of an intact rope. The analyses are focused on homogeneous small-scale ropes with nonlinear constitutive laws and overall diameters equal to 6 mm that are damaged at ropes midspan location. Ropes with a wide range of degree of damage (percentage of the broken components of the damaged cross-section with respect to the intact rope) and degree of asymmetry (change of the center of stiffness of the damaged cross-section with respect to the intact rope) are considered in this study. Preliminary results given by the proposed model are found to be in good agreement with available static tension tests on asymmetrically damaged polyester ropes and 3D finite element simulations with lower computational cost. Compared to the solutions obtained by previous analytical models reported in the literature, the range of applicability associated to the degree of damage to rope cross-section (number of broken rope components) is extended.
1 INTRODUCTION

Ropes are employed in many engineering applications including cranes, lifts, mine hoisting, bridges, cableways, electrical conductors, offshore mooring systems and so on\cite{1,2}. Mechanical demands, abrasion, and environmental interaction (corrosion; ultra-violet light, chemical, and heat exposures; etc) degrade the properties of the individual rope components continuously during rope operational service. This degradation process could result in the complete rupture of some rope components and eventually will lead to rope failure. The service life of a rope can be greatly extended by following a planned program of installation, operation, maintenance, and inspection\cite{3}. Hence, damage-tolerance property (i.e., the ability of a rope to withstand damage), is an essential parameter for rope design, rope evaluation during operational service, and for developing discard criteria according to rope usage based on the residual strength and deformation capacity that the damaged rope can sustain.

Based on several experimental and theoretical studies\cite{4-9}, the impact of the presence of broken rope components on overall rope response (stiffness, residual strength and deformation capacity) depends on the length of the rope, number of broken rope components (degree of damage) and their distributions throughout rope cross-section (symmetric and asymmetric) and along the rope length, and rope construction. These studies conclude that the effect of rope components breaks on rope response is not always directly proportional to the equivalent loss of cross-sectional area (the so-called net area effect) and asymmetric damage distribution induces a lateral displacement of the rope and a non-uniform strain (and stress) distribution throughout rope cross-section, resulting in premature failure (deformation capacity) of the asymmetrically damaged rope relative to the intact rope.

In this paper a simplified nonlinear mechanical model is proposed to estimate the residual strength and deformation capacity of ropes with asymmetric damage distribution. In the proposed model, the damaged rope is assumed to behave as a nonlinear beam under biaxial bending and axial load with Bernoulli’s kinematic hypothesis. In this study, damage corresponds to the complete rupture of one or more rope components that belong to the outermost layer of a particular rope cross-section. Preliminary results given by the proposed model are validated by comparisons with available with available static tension tests on asymmetrically damaged polyester ropes\cite{10}, 3D finite element simulations\cite{9} and results given by previous analytical models\cite{7,8}.

2 PROPOSED MECHANICAL MODEL

It is assumed that the response of an asymmetric damaged rope can be accurately predicted by uncoupled biaxial rope responses (i.e., two uniaxial bending analyses). Biaxial bending arises from the unbalanced radial contact forces within rope cross-section, which are related to the initial helical geometry configuration of the rope components, due to the asymmetric damage distribution as explained in a previous study\cite{8}. Therefore, the proposed model is based on the equations that govern the response of a nonlinear planar beam considering Bernoulli’s kinematic hypothesis (plane sections of a beam remain plane and normal to its centroidal axis after deformation)\cite{11} and assumptions adopted based on the results given by 3D FE simulations\cite{9}.
From the balance of forces of a planar ($xy$ plane) rope segment $[x, x+\Delta x]$ in the axial and transverse directions, the moment equilibrium about the point $P$, and considering the limit as $\Delta x \to 0$, the following equilibrium equations are obtained (Figure 1)\cite{11}.

\[
\frac{dH}{dx} + h(x) = 0 \tag{1}
\]

\[
\frac{dV_y}{dx} + q_y(x) = 0 \tag{2}
\]

\[
\frac{dM_{zz}}{dx} + V_y \left(1 + \frac{du}{dx}\right) - H \frac{dv}{dx} + m_z(x) = 0 \tag{3}
\]

where $u(x)$ is the displacement of the centroid in the axial direction, $v(x)$ is the displacement of the centroid in the transverse direction (Figure 1); $H(x)$ and $V_y(x)$ are the horizontal and vertical forces relative to the axial and transverse direction of the undeformed rope; $M_{zz}(x)$ is the bending moment around $z$ axis; and $h(\xi)$, $q_y(\xi)$, and $m_z(\xi)$ are distributed axial force, transverse force and bending moment respectively along the rope segment.

Considering that the slopes of the deformed longitudinal axis of the rope (beam) are small compared to unity, small angle approximation assumption is used in this study. Hence, rope curvature can be approximated by the second derivative of the lateral displacement $v(x)$ (respect to the initial rope configuration). Using the Bernoulli’s kinematic hypothesis and the secant formulation to estimate the bending stiffness in order to establish the moment-curvature relationship for a rope\cite{12}; the fact that axial strain is constant along rope components\cite{9} which implies that $d^2u(x)/dx^2 = 0$, and the absence of distributed axial force and bending moment; lead to the following fourth order differential equation after combining the first derivate of Eq. (3) to Eq. (2)

\[
(EI_{zz})_{sec} \frac{d^4v}{dx^4} - q_y(x) \left(1 + \frac{du}{dx}\right) - H(x) \frac{d^2v}{dx^2} = 0 \tag{4}
\]
\[
(\text{EI}_{\text{yy}})_{\text{sec}} \frac{d^4 w}{dx^4} - q_z(x) \left(1 + \frac{du(x)}{dx}\right) - H(x) \frac{d^2 w}{dx^2} = 0
\]

where \(w(x)\) is the rope deflection in the \(xz\) plane, \(q_z(x)\) is the distributed force along the rope length in the \(z\) direction and \((\text{EI}_{zz})_{\text{sec}}\) and \((\text{EI}_{yy})_{\text{sec}}\) are the secant bending stiffness of the cable respect to the \(z\)-axis and \(y\)-axis respectively. The Eq. (5) is obtained using the same kinematic assumptions considered in Eq (4) and the balance of forces applied for the \(xz\) plane in a similar way accounted for the \(xy\) plane (Figure 1).

The values of the distributed forces \(q_y(x)\) and \(q_z(x)\) can be estimated on the same assumption established in [8]: radial forces exerted on rope core are not in equilibrium due to the loss of symmetry of the rope cross-section. For the case of a two-layer rope, radial forces exerted on the core depends on the values of the axial force \((T)\) and curvature \((\kappa)\) of the components of the second layer. Thus, in order to illustrate the process to estimate the values of \(q_y(x)\) and \(q_z(x)\) assuming that the unbroken rope components are mainly in radial contact with the rope core, considered the damaged cross-section shown in Figure 2 in which component 6 is broken: thus the radial line body force rope component 3 is unbalanced due to the failure of rope component 6 in the \(y_1\) direction. Hence, the net line transverse forces along the longitudinal axis of the rope in both \(xy\) and \(xz\) planes are given by

\[
q_y = -\frac{\kappa \cdot T_2 \cos \varphi}{\cos \theta_2}
\]

\[
q_z = -\frac{\kappa \cdot T_2 \sin \varphi}{\cos \theta_2}
\]

where \(\theta_2\) is the helix angle of the components of the second layer and \(\varphi\) is the swept angle of the rope components.

![Figure 2. (a) Radial equilibrium; (b) boundary conditions](image)

The boundary conditions specified to the rope model considered that at \(x = 0\) rope cross-section is fully clamped: \(v_{x=0} = 0; \ w_{x=0} = 0; \ v'_{x=0} = 0; \) and \(w'_{x=0} = 0; \) and at \(x = L\) (rope length)
an axial displacement history is specified \((u_{x=L} = \Delta u)\) and the cross-section is prevented from rotating and laterally deflecting: \(v_{x=L} = 0;\) \(v_{x=L} = 0;\) \(v_{x=L} = 0;\) and \(w_{x=L} = 0\) where the notation \((\cdot)' = d(\cdot)/dx\) is used for the first derivate (Figure 2b). The general closed-form solution for \(v(x)\) (Eq. 4) and \(w(x)\) (Eq. 5) are given by

\[
v(x) = \frac{a_y}{b_y} e^{\frac{\sqrt{b_y} x}{b_y}} + C_5 e^{\frac{-\sqrt{b_y} x}{b_y}} + C_6 \frac{e^{\sqrt{b_y} x}}{b_y} + C_7 x + C_8
\]

\[
w(x) = \frac{a_x}{b_x} e^{\frac{\sqrt{b_x} x}{b_x}} + C_4 e^{\frac{-\sqrt{b_x} x}{b_x}} + C_3 \frac{e^{-\sqrt{b_x} x}}{b_x} + C_9 x + C_{10}
\]

where \(a_x(\varepsilon) = -\kappa_2 T_2 / \cos \theta_2 (1 + \Delta u / L) / (EI_{zz})_{sec};\) \(a_y(\varepsilon) = -\kappa_2 T_2 / \cos \theta_2 (1 + \Delta u / L) / (EI_{yy})_{sec};\) \(b_x(\varepsilon) = \kappa_2 T_2 / \cos \theta_2 (EI_{zz})_{sec}; \) \(b_y(\varepsilon) = \kappa_2 T_2 / \cos \theta_2 / (EI_{yy})_{sec}; \) \(c = 2\pi / p;\) and the constants \(C_i\) are given by the boundary conditions of the problem. If \((errH)_{i+1} > tol,\) after Step 8 of the algorithm, the expression to compute \((E_{sec})_{i+1}\) is given by

\[
(E_{sec})_{i+1} = \sum_k \left[ (AE(e_k))_k \right]_{i+1}^{j+1} / \sum_k A_k
\]

which is necessary to solve the Eqs. (4) and (5) accounting for the rope deformed configuration in the \((i+1)\) iteration during the \(j\)th analysis step. The same expression, considering a small value of the axial deformation value \(e_k\) for the case of straight rope configuration, can be used to evaluate lateral rope deflection (Eqs. 8 and 9) for the first iteration \((i = 0; i \rightarrow i+1)\) in the \(j\)th analysis step (Step 3).

3 ANALYSIS OF NUMERICAL EXAMPLES

The validation of the proposed model is presented by comparisons between rope axial load and rope axial strain curves obtained from 3D FE simulations\(^9\), NLBM (proposed model), two linear models (Lanteigne and MacDougall and Bartlett (MB)) reported in the literature\(^7,8\) and experimental data of asymmetrically damaged small polyester ropes\(^10\). Polyester ropes with the same type of construction but different degree of damage and degree of asymmetry \(IA\) (change of the center of stiffness of the damaged cross-section with respect to the intact rope)\(^9\) are used in the validation process. Ropes have a diameter equal to 6 mm and they are comprised of eight helical components (second layer of the rope) wound around a straight core (first layer of the rope), with initial pitch distance \(p_0\) and helix angle \(\theta_0\) (angle between the longitudinal axis of the rope and local axis of the rope components) equal to 81 mm and 9.5º respectively. The indices used to identify a particular cross-section is as follows: \(R_{ij}\) where \(R\) refers to the cross-section and the indices \(i\) and \(j\) refer to the broken rope components and the unbroken components numbered \(m\) and \(n\) of the rope \(R_{ij}\) are cited in the text as \(R_{ij,mn}\) \((i,j,m,\text{and } n\text{ vary from 0 to 9})\) based on the undamaged rope cross-section \((R_u)\) shown in the corresponding figures. The undamaged (intact) rope has a strength and deformation capacity values equal to 23.5 kN and 0.117 (natural strain) respectively.
Table 1: Iterative cross-sectional displacement control algorithm

**Step 1:** Given a rope axial displacement increment value $\Delta u_j$, an initial horizontal (axial) force value $H_j^i$ ($i = 0$) is computed considering a straight rope. The centroid of the $k$th unbroken rope component is considered as the generic point $t$ and $A_k$ is the its cross-sectional area. The normal stress $\sigma_k$ is computed from its constitutive law. $u_j = u_j + \Delta u_j$ and $H_j^i = \sum A_k \sigma_k (u_j/L_0) \cos^3(\theta_k)$.

**Step 2:** Compute the net transverse forces $(q_y)_j^i(x, \varepsilon_j)$ and $(q_z)_j^i(x, \varepsilon_j)$ through Eqs. (6) and (7) respectively, considering the corresponding cross-section damage distribution and the initial helical geometry of the rope components.

**Step 3:** Determine the lateral deflections (i.e. the updated rope configuration $(i \rightarrow i+1)$) $v_{j+1}^i(x)$ and $w_{j+1}^i(x)$ using Eqs. (9) and (10) for the total rope axial displacement given in the $j$th incremental step of the analysis $u_j$ (Step 1).

**Step 4:** The displacement field $u(x)_{j+1}$, $v(x)_{j+1}$ and $w(x)_{j+1}$ is computed for the $k$th unbroken rope component considering the centroid of each unbroken rope component as the generic point (coordinates $(z_k, y_k)$): $u_x = u - y (v')$, $u_y = v (x)$, $u_z = w (x)$.

**Step 5:** Compute the axial strain value $(\varepsilon_k)_{j+1}$ for the $k$th unbroken rope component considering its centroid as the generic point: $\varepsilon_k(x) = du/k$. The updated value of the helix angle $\theta_k$: $\tan(\theta_k)_{j+1} = 2\pi R_{hk}^{j+1} / p(\varepsilon_k)$, where $(R_{hk})_{j+1}$ is the equivalent helix radius of the $k$th unbroken rope component in its deformed configuration for the iteration $(i+1)$ (distance from the centroid of the rope core to the centroid of the rope component) and $p(\varepsilon)$ is the updated pitch.

**Step 6:** The normal stress for the $k$th unbroken rope component $(\sigma_k)_{j+1}$ is calculated using its constitutive law considering its centroid as the generic point. The axial force developed by the $k$th unbroken rope component is given by $(T_k)_{j+1} = A_k (\sigma_k)_{j+1}$.

**Step 7:** Compute the updated horizontal force value $H_j^{i+1}$ as follows $H_j^{i+1} = \sum (T_k)^{i+1} \cos(\theta_k)^{i+1}$.

**Step 8:** Compute the error value of the horizontal force as $(errH)_j^{i+1} = |H_j^{i+1} - H_j^i| / H_j^i$.

If $(errH)_j^{i+1} > tol$ then **Step 9:** $H_j = H_j^{i+1}$.

If $(errH)_j^{i+1} \leq tol$ then stop.

---

**Table 1: Iterative cross-sectional displacement control algorithm**

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Given a rope axial displacement increment value $\Delta u_j$, an initial horizontal (axial) force value $H_j^i$ ($i = 0$) is computed considering a straight rope. The centroid of the $k$th unbroken rope component is considered as the generic point $t$ and $A_k$ is the its cross-sectional area. The normal stress $\sigma_k$ is computed from its constitutive law. $u_j = u_j + \Delta u_j$ and $H_j^i = \sum A_k \sigma_k (u_j/L_0) \cos^3(\theta_k)$.</td>
</tr>
<tr>
<td>2.</td>
<td>Compute the net transverse forces $(q_y)_j^i(x, \varepsilon_j)$ and $(q_z)_j^i(x, \varepsilon_j)$ through Eqs. (6) and (7) respectively, considering the corresponding cross-section damage distribution and the initial helical geometry of the rope components.</td>
</tr>
<tr>
<td>3.</td>
<td>Determine the lateral deflections (i.e. the updated rope configuration $(i \rightarrow i+1)$) $v_{j+1}^i(x)$ and $w_{j+1}^i(x)$ using Eqs. (9) and (10) for the total rope axial displacement given in the $j$th incremental step of the analysis $u_j$ (Step 1).</td>
</tr>
<tr>
<td>4.</td>
<td>The displacement field $u(x)<em>{j+1}$, $v(x)</em>{j+1}$ and $w(x)_{j+1}$ is computed for the $k$th unbroken rope component considering the centroid of each unbroken rope component as the generic point (coordinates $(z_k, y_k)$): $u_x = u - y (v')$, $u_y = v (x)$, $u_z = w (x)$.</td>
</tr>
<tr>
<td>5.</td>
<td>Compute the axial strain value $(\varepsilon_k)<em>{j+1}$ for the $k$th unbroken rope component considering its centroid as the generic point: $\varepsilon_k(x) = du/k$. The updated value of the helix angle $\theta_k$: $\tan(\theta_k)</em>{j+1} = 2\pi R_{hk}^{j+1} / p(\varepsilon_k)$, where $(R_{hk})_{j+1}$ is the equivalent helix radius of the $k$th unbroken rope component in its deformed configuration for the iteration $(i+1)$ (distance from the centroid of the rope core to the centroid of the rope component) and $p(\varepsilon)$ is the updated pitch.</td>
</tr>
<tr>
<td>6.</td>
<td>The normal stress for the $k$th unbroken rope component $(\sigma_k)<em>{j+1}$ is calculated using its constitutive law considering its centroid as the generic point. The axial force developed by the $k$th unbroken rope component is given by $(T_k)</em>{j+1} = A_k (\sigma_k)_{j+1}$.</td>
</tr>
<tr>
<td>7.</td>
<td>Compute the updated horizontal force value $H_j^{i+1}$ as follows $H_j^{i+1} = \sum (T_k)^{i+1} \cos(\theta_k)^{i+1}$.</td>
</tr>
<tr>
<td>8.</td>
<td>Compute the error value of the horizontal force as $(errH)_j^{i+1} =</td>
</tr>
<tr>
<td>9.</td>
<td>If $(errH)_j^{i+1} &gt; tol$ then $H_j = H_j^{i+1}$. If $(errH)_j^{i+1} \leq tol$ then stop.</td>
</tr>
</tbody>
</table>
Tensile stress distributions throughout asymmetric damaged rope cross-section and capacity curves predicted by the numerical and analytical models are compared in Figure 3 ((a)-(d)), in which the rope R₁₂ is selected to carry out these comparative analyses. This rope has an index of asymmetry \( I_A \) equal to 0.31\[^9\]. As the 3D FE models predict a gradual variation in the strain/stress distribution throughout unbroken rope components cross-sections, the maximum and minimum values are plotted for comparison purposes and also to define a range of values that allow determining the accuracy of the estimated values given by the analytical models (hereafter referred as admissible values). The stress gradient throughout damaged rope cross-section estimated by the Lanteigne and MB models is higher than the gradient estimated by the 3D FE models: they overestimate tensile stress values of rope components that are adjacent to damage in about 20%; predicted tensile stress values of rope components close to the rope centroid are within the range values defined by the 3D FE simulations; and underestimate stress values of the components that are on the opposite side to damage in about 15% and over 100% for components R₁₂,82 and R₁₂,65, respectively. Conversely to the results provided by Lanteigne and M-B models, the unbroken components tensile stress values estimated by the NLBM lay in the ranges of the admissible values defined the 3D FE model, matching quite well with the upper values of the aforementioned ranges for the entire range of the rope axial strain. It is important to point out that additional numerical simulations in the same type of rope construction with \( I_A \) values in the range of 0.093 and 0.54 in which the same previous comparison pattern between the unbroken rope component tensile stress values given by the 3D FE and NLBM models is found\[^13\]. In this recently published paper, it is shown that as the value of \( I_A \) gets bigger rope components on the opposite side to damage start developing compressive load resulting in a more flexible damaged rope response relative to 3D FE and NLBM simulations.

Regarding the predicted capacity curve obtained from the NLBM (Figure 3e), it is slightly stiffer than the curves predicted by Lanteigne, M-B and 3D FE models, it is bounded by the experimental curves although Exp. 1 curve fails earlier probably due to a progressive damage of rope components around rope terminations as reported by the researcher who conducted the tests\[^10\]. NLBM curves overestimates the residual strength and deformation capacity of the rope in about 1% and 8% respectively when compared with values provided by the 3D FE simulation. Curves predicted by Lanteigne and MB models compare quite well between each other, they are also bounded by the experimental curves and as the rope axial strain value increases both curves approaches to their lower bound (Exp. 2 curve). As these two curves get more flexible for increasing values of the rope axial strain, due to the large curvatures and large lateral deflection of the ropes estimated by the models, they underestimate the residual strength and deformation capacity of the rope in about 30% and 15% respectively.

A summary of the residual strength and deformation capacity values and their comparisons relative to the values given by the 3D FE simulation are presented in Table 2 for all the models considered in this study for a wide range of \( I_A \) values (0.093-0.54). NLBM simulations provide better estimates for a broader degree of asymmetry values (\( I_A \) values) when compared to the results extracted from a traditional and well accepted 3D FE approach. The NLBM simulations overestimate in a range of [8%, 12%] the residual strength, in which the lower bound is related to the comparison of the W₁₃₆ rope and the upper bound to the W₁₂₃₄ rope; and overestimate as well in less than 2% the failure axial strain relative to 3D FE models results. Conversely, Lanteigne and MB models provide satisfactory estimates of the
residual strength and deformation capacity (values are underestimated in about 10%) for ropes with low $IA$ values (less than 0.16).

Figure 3. (a)-(d) Tensile stress distribution; (e) Capacity curves of rope $R_{12}$
It is important to point out that in Figure 3e has also been plotted the damaged capacity curve associated to the net area effect model. This model simply neglects the contribution to rope response of the broken rope components and the intact rope strength is multiplied by the parameter $\rho_A$ that represents the remaining cross-sectional area of the rope and the deformation capacity is the same as the intact rope. For this particular rope, the residual strength and deformation capacity are overestimated in 3% and 5% respectively by this model relative to the 3D FE simulation values. In a more extensive analysis, it is concluded that for damaged ropes with $IA$ values less than 0.44 rope residual strength can be satisfactorily estimated based upon the net area effect model and the deformation capacity is overestimated in less than 7%\[13\].

<table>
<thead>
<tr>
<th>Rope</th>
<th>Model</th>
<th>Deformation capacity</th>
<th>Residual strength (kN)</th>
<th>Diff (%) Failure strain</th>
<th>Diff (%) Residual strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>W(136) ($IA = 0.093$)</td>
<td>3D FE</td>
<td>0.116</td>
<td>15.60</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td></td>
<td>Lanteigne</td>
<td>0.104</td>
<td>14.26</td>
<td>-10.3%</td>
<td>-8.6%</td>
</tr>
<tr>
<td></td>
<td>MB</td>
<td>0.103</td>
<td>14.20</td>
<td>-11.2%</td>
<td>-9.0%</td>
</tr>
<tr>
<td></td>
<td>NLBM</td>
<td>0.115</td>
<td>16.65</td>
<td>-0.86%</td>
<td>6.7%</td>
</tr>
<tr>
<td></td>
<td>Exp. 1</td>
<td>0.114</td>
<td>17.03</td>
<td>-1.72%</td>
<td>9.2%</td>
</tr>
<tr>
<td></td>
<td>Exp. 2</td>
<td>0.10</td>
<td>13.39</td>
<td>-13.8%</td>
<td>-14.2%</td>
</tr>
<tr>
<td></td>
<td>Exp. 3</td>
<td>0.096</td>
<td>12.14</td>
<td>-17.2%</td>
<td>-22.2%</td>
</tr>
<tr>
<td></td>
<td>Net area</td>
<td>0.117</td>
<td>15.6</td>
<td>0.86%</td>
<td>0.0%</td>
</tr>
<tr>
<td>W(1) ($IA = 0.16$)</td>
<td>3D FE</td>
<td>0.114</td>
<td>20.50</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td></td>
<td>Lanteigne</td>
<td>0.102</td>
<td>18.21</td>
<td>-10.2%</td>
<td>-11.2%</td>
</tr>
<tr>
<td></td>
<td>MB</td>
<td>0.103</td>
<td>18.20</td>
<td>-9.4%</td>
<td>-11.2%</td>
</tr>
<tr>
<td></td>
<td>NLBM</td>
<td>0.115</td>
<td>22.15</td>
<td>0.8%</td>
<td>8.0%</td>
</tr>
<tr>
<td></td>
<td>Exp. 1</td>
<td>0.117</td>
<td>19.83</td>
<td>2.3%</td>
<td>-3.3%</td>
</tr>
<tr>
<td></td>
<td>Exp. 2</td>
<td>0.117</td>
<td>20.03</td>
<td>2.3%</td>
<td>-2.3%</td>
</tr>
<tr>
<td></td>
<td>Net area</td>
<td>0.117</td>
<td>20.9</td>
<td>2.6%</td>
<td>2%</td>
</tr>
<tr>
<td>W(12) ($IA = 0.31$)</td>
<td>3D FE</td>
<td>0.114</td>
<td>17.83</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td></td>
<td>Lanteigne</td>
<td>0.10</td>
<td>12.97</td>
<td>-11.8%</td>
<td>-27.3%</td>
</tr>
<tr>
<td></td>
<td>MB</td>
<td>0.097</td>
<td>12.0</td>
<td>-15%</td>
<td>-32.7%</td>
</tr>
<tr>
<td></td>
<td>NLBM</td>
<td>0.115</td>
<td>19.35</td>
<td>0.8%</td>
<td>8.5%</td>
</tr>
<tr>
<td></td>
<td>Exp. 1</td>
<td>0.076</td>
<td>11.54</td>
<td>-33.4%</td>
<td>-35.3%</td>
</tr>
<tr>
<td></td>
<td>Exp. 2</td>
<td>0.117</td>
<td>18.61</td>
<td>2.6%</td>
<td>4.4%</td>
</tr>
<tr>
<td></td>
<td>Net area</td>
<td>0.117</td>
<td>18.3</td>
<td>2.6%</td>
<td>2.5%</td>
</tr>
<tr>
<td>W(123) ($IA = 0.44$)</td>
<td>3D FE</td>
<td>0.111</td>
<td>14.61</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td></td>
<td>NLBM</td>
<td>0.111</td>
<td>15.97</td>
<td>0.0%</td>
<td>9.3%</td>
</tr>
<tr>
<td></td>
<td>Exp. 1</td>
<td>0.084</td>
<td>9.72</td>
<td>-24.4%</td>
<td>-33.5%</td>
</tr>
<tr>
<td></td>
<td>Net area</td>
<td>0.117</td>
<td>15.7</td>
<td>4.9%</td>
<td>7.23%</td>
</tr>
<tr>
<td>W(1234) ($IA = 0.54$)</td>
<td>3D FE</td>
<td>0.11</td>
<td>11.87</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td></td>
<td>NLBM</td>
<td>0.11</td>
<td>13.03</td>
<td>0.0%</td>
<td>9.8%</td>
</tr>
<tr>
<td></td>
<td>Exp. 1</td>
<td>0.11</td>
<td>10.14</td>
<td>0.0%</td>
<td>-14.6%</td>
</tr>
<tr>
<td></td>
<td>Net area</td>
<td>0.117</td>
<td>13.1</td>
<td>6.4%</td>
<td>10%</td>
</tr>
</tbody>
</table>
4 CONCLUSIONS

A simplified nonlinear mechanical model is proposed to estimate the residual strength and deformation capacity of ropes with asymmetric damage distribution. The proposed model is validated by comparisons with 3D FE simulations, experimental data and two linear models reported in the literature. Numerical simulations show that an asymmetrically damaged rope subjected to tensile load laterally deflects inducing additional bending stresses in unbroken rope components. The proposed model (NLM) is applicable with satisfactory results, relative to the results provided by 3D FE simulations and experimental data, to damaged ropes with a wide range of IA values (0.093-0.54); thus, the range of applicability associated to the degree of damage and degree of asymmetry is extended relative to the two linear models previously reported in the literature considered in this study. The NLBM and 3D FE models, however, suggest that for IA values less than 0.44, rope residual strength can be satisfactory estimated based upon the net area effect model.

Based on the ropes analyzed in this study, the computational efficiency of the proposed algorithm is very high relative to the results given by the 3D FE nonlinear simulations. The robustness of the present algorithm is confirmed by the fact that all the capacity curves and deformed configurations associated to the damaged ropes analyzed, were accurately estimated and they converged after few iterations (less or equal than five). Further comparisons, however, with experimental data of bigger ropes (the iterative algorithm can be readily extended for the case of asymmetrically damaged multilayered ropes) are needed to verify the assumptions made and establish the range of applicability of the proposed model (NLBM).

REFERENCES


COMPARATIVE ANALYSIS OF NEW SHEAR LOCKING-FREE FINITE ELEMENT WITH OTHER COMMONLY USED APPROACHES IN VIBRATION ANALYSIS OF MINDLIN PLATES

I. SENJANOVIĆ¹, N. VLADIMIR¹, D.S. CHO², M. JOKIĆ¹, T.M. CHOI³

¹University of Zagreb, Faculty of Mechanical Engineering and Naval Architecture
Ivana Lucica 5, 10000 Zagreb, Croatia
nikola.vladimir@fsb.hr, www.fsb.unizg.hr

²Pusan National University, Department of Naval Architecture and Ocean Engineering
63 beon-gil 2, Busandaehak-ro, Geumjeong-gu, Busan, 609-735, Korea

³Createch Co., Ltd.
Room 1312, Centum IS tower, 1209 Jaesong-dong, Haeundae-gu, Busan, 612-050, Korea

Key words: Mindlin Plate Theory, Shear Locking, Vibration, FEM.

Abstract. This paper is dedicated to validation of newly developed shear locking-free finite element (FE) for the vibration analysis of Mindlin plates. First, the state-of-the-art in thick plate vibration theories and analysis methods is described and basic equations of the original and advanced Mindlin plate theories, respectively, are listed. Then, a detailed description of shear-locking free rectangular finite element is provided. Bending deflection is used as a potential function for the definition of total (bending and shear) deflection and angles of cross-section rotations. Extensive calculations are done by using developed in-house codes and standard commercial FE tools. Also, comparisons with analytical solutions and energy-based assumed mode method results are included. In addition, convergence of the developed finite element is checked. Natural vibration analyses of rectangular plates having different edge constraints, utilizing the proposed quadrilateral FE, show very good agreement with other commonly used methods.

1 INTRODUCTION

Thick plates are primary constitutive members in many engineering structures. They can be used as engine foundations, elements of reinforced concrete bridges or parts of different floating structures, etc. The thick plate theory represents an issue for very long time, i.e. from the first works published by Reissner and Mindlin [1,2]. As can be seen in the literature survey presented by Liew et al. [3], many concepts for the vibration analysis of thick plates, based both on analytical and numerical solution of equilibrium equations, have been worked out. There are different analytical methods differing on which functions are kept as fundamental ones in the reduction of the system of differential equations of motion [4]. However, analytical solutions are applicable only to simply supported plates or plates with two opposite edges simply supported. For plates with arbitrary edge constraints, including also elastically restrained edges, different variants of the Rayleigh-Ritz (energy) method are on disposal. Their accuracy is dependent on
the chosen set of orthogonal functions for the assumed natural modes, where two dimensional polynomials or static Timoshenko beam deflection functions, [5] and [6,7,8,9], respectively, can be used. Nowadays, FEM represents advanced and widespread numerical tool in structural engineering. It can be applied to vibration analysis of thick plates of any shape including all possible sets of boundary conditions. However, shear locking problem associated with finite elements can appear. Namely, in the Mindlin thick plate theory shear deformations are taken into account, and application of ordinary low-order finite element is not capable to reproduce the pure bending modes in the limit case of thin plate. This shear locking problem arises due to inadequate dependence among transverse deflection and two rotations. Usually, the developed finite elements are based on direct application of the Mindlin plate theory, which deals with plate deflection and angles of rotation as three basic variables. Although, there are different methods for shear locking elimination, it seems to be more reasonable to avoid its appearance implicitly.

This paper outlines a new procedure for determining properties of thick plate finite elements, based on the modified Mindlin theory for moderately thick plate. Bending deflection is used as a potential function for the definition of total (bending and shear) deflection and angles of cross-section rotations [4,10]. As a result of the introduced interdependence among displacements, the shear locking problem, present and solved in known finite element formulations, is implicitly avoided [11,12]. Based on the developed theoretical model, in-house code has been developed and applied to vibration analyses of rectangular plates with different thickness to length (width) ratios and different sets of boundary conditions. Since the special attention is paid to the validation of newly introduced finite element, comparisons are made to the results available in the relevant literature, analytical solutions [4,10], solutions obtained by the assumed mode method [8] as well as to those obtained by general purpose finite element software [13,14]. In addition, convergence study is performed and shows superiority of the developed FE in the case of thin and moderately thick plate.

2 FUNDAMENTALS OF THE ORIGINAL AND ADVANCED MINDLIN THEORIES

The Mindlin theory deals with three general displacements, i.e. plate deflection \( w \), and angles of cross-section rotation about \( y \) and \( x \) axis, \( \psi_x \) and \( \psi_y \), respectively. The following relations between sectional forces, i.e. bending moments, \( M_x \) and \( M_y \), torsional moments, \( M_{xy} \) and \( M_{yx} \), and transverse shear forces, \( Q_x \) and \( Q_y \), and displacements via deformations are specified, [1,2]

\[
M_x = D \left( \frac{\partial \psi_y}{\partial x} + \nu \frac{\partial \psi_x}{\partial y} \right), \quad M_y = D \left( \frac{\partial \psi_x}{\partial y} + \nu \frac{\partial \psi_y}{\partial x} \right), \quad M_{xy} = M_{yx} = \frac{1}{2}(1-\nu)D \left( \frac{\partial \psi_x}{\partial y} + \frac{\partial \psi_y}{\partial x} \right),
\]

\[
Q_x = S \left( \frac{\partial w}{\partial x} + \psi_y \right), \quad Q_y = S \left( \frac{\partial w}{\partial y} + \psi_x \right),
\]

where

\[
D = \frac{Eh^3}{12(1-\nu^2)}, \quad S = kGh,
\]
is plate flexural rigidity and shear rigidity, respectively, \( h \) is plate thickness, \( k \) is shear coefficient, \( E \) and \( G = E/(2(1+v)) \) are Young’s and shear modulus, respectively, while \( v \) is Poisson’s ratio.

The plate is loaded with transverse inertia load and distributed inertia moments

\[
q = -\bar{m} \frac{\partial^2 w}{\partial t^2}, \quad \bar{m}_x = \int \frac{\partial^3 \psi_x}{\partial t^2}, \quad \bar{m}_y = \int \frac{\partial^3 \psi_y}{\partial t^2}
\]

where \( \bar{m} = \rho h \) and \( J = \rho h^3 / 12 \) are plate specific mass per unit area and its moment of inertia, respectively, and \( \rho \) is mass density. Equilibrium of sectional and inertia forces, i.e. moment about \( y \) and \( x \) axis and transverse forces read

\[
\frac{\partial M_x}{\partial x} + \frac{\partial M_y}{\partial y} - Q_x = \bar{m}_x, \quad \frac{\partial M_y}{\partial x} + \frac{\partial M_x}{\partial y} - Q_y = \bar{m}_y, \quad \frac{\partial Q_x}{\partial x} + \frac{\partial Q_y}{\partial y} = -q.
\]

By substituting Eqs. (1) and (3) into (4) one arrives at three differential equations of motion (well-known Mindlin equations)

\[
\frac{D}{S} \left[ \frac{\partial^2 \psi_x}{\partial x^2} + \frac{1}{2} (1-v) \frac{\partial^2 \psi_y}{\partial y^2} + \frac{1}{2} (1+v) \frac{\partial^2 \psi_y}{\partial x \partial y} \right] - \left( \frac{\partial w}{\partial x} + \psi_x \right) - \frac{J}{S} \frac{\partial^2 \psi_x}{\partial t^2} = 0,
\]

\[
\frac{D}{S} \left[ \frac{\partial^2 \psi_y}{\partial y^2} + \frac{1}{2} (1-v) \frac{\partial^2 \psi_x}{\partial x^2} + \frac{1}{2} (1+v) \frac{\partial^2 \psi_x}{\partial x \partial y} \right] - \left( \frac{\partial w}{\partial y} + \psi_y \right) - \frac{J}{S} \frac{\partial^2 \psi_y}{\partial t^2} = 0,
\]

\[
\Delta w + \frac{\partial \psi_x}{\partial x} + \frac{\partial \psi_y}{\partial y} + \frac{\bar{m}}{S} \frac{\partial^2 w}{\partial t^2} = 0,
\]

where \( \Delta() = \frac{\partial^2()}{\partial x^2} + \frac{\partial^2()}{\partial y^2} \) is the Laplace differential operator.

According to the advanced vibration theory of moderately thick plate [10] deflection of plate consists of beam deflection and shear contribution as in Timoshenko beam theory [15], Fig. 1

\[
w(x,y,t) = w_b(x,y,t) + w_s(x,y,t).
\]

It is also specified that only bending deflection causes the rotation of plate cross-section, Fig. 1

\[
\psi_x = \frac{\partial w_b}{\partial x}, \quad \psi_y = -\frac{\partial w_b}{\partial y}.
\]

Moreover, according to advanced vibration theory of moderately thick plate [4], system of three differential equations (5), (6) and (7), respectively, can be reduced to only one with the unknown bending deflection

\[
D \Delta w_b - J \left( 1 + \frac{\bar{m} D}{JS} \right) \frac{\partial^2 w_b}{\partial t^2} + \bar{m} \frac{\partial^2 w_b}{\partial t^2} \left( w_b + J \frac{\partial^2 w_b}{\partial t^2} \right) = 0
\]

The bending deflection is a potential function since the remaining displacements \( w_x, \psi_x \) and \( \psi_y \) can be expressed by its derivatives. Once \( w_b \) is determined, the total deflection reads
Analytical solutions for simply supported plates and plates with two opposite edges simply supported are presented in [4]. Rigorous solutions for complex boundary conditions can be obtained by employing the method of separation of variables, as shown in [10].

3 SHEAR LOCKING-FREE FINITE ELEMENT

Complete derivation of rectangular and triangular shear locking-free finite elements is presented in [12] and here theoretical background related to the former is reviewed. A general finite element with \( n \) nodes and three d.o.f. per node, i.e. deflection and rotations around the \( x \) and \( y \) axis, is considered. The ordinary procedure for determining stiffness and the mass matrix in the case of a thin plate is used [16]. Bending deflection as a potential function is assumed in a polynomial form with a number of unknown coefficients which corresponds to the total number of d.o.f. \( N=3n \)

\[
\{w\} = \{a\} \{P\}_h ,
\]

where \( \{a\} \) is a row vector with terms \( a_i, i=0,1,...,N-1 \), and

\[
\{P\}_h = \{P\}_h = \{1, x, y, x^2, xy, y^2 \}.
\]

Total static deflection can be expressed in the following form

\[
w = \{a\} \left( \{P\}_h - \frac{D}{S} \frac{\partial^2 \{P\}_h}{\partial x^2} - \frac{D}{S} \frac{\partial^2 \{P\}_h}{\partial y^2} \right)
\]

Angles of rotation (9) yield

\[
\psi_x = -\{a\} \frac{\partial \{P\}_h}{\partial x}, \quad \psi_y = -\{a\} \frac{\partial \{P\}_h}{\partial y}.
\]

By taking coordinate values \( x_l \) and \( y_l \) for each node, \( l=1,2...,n \), into account in Eqs. (14) and (15), the relation between the nodal displacements and the unknown coefficients \( a_i \) is obtained

\[
\{\delta\} = [C]\{a\},
\]
where \([C]\) includes \(x_i\) and \(y_i\) and

\[
\{ \delta \} = \begin{bmatrix}
\{ \delta \}_1 \\
\vdots \\
\{ \delta \}_n
\end{bmatrix}, \quad \{ \delta \}_i = \begin{bmatrix}
w_i \\
\phi_i
\end{bmatrix}.
\] (17)

Now, for the given nodal displacement vector \(\{ \delta \}\), the corresponding coefficient vector \(\{ a \}\) can be determined from (16)

\[
\{ a \} = [C]^{-1}\{ \delta \}.
\] (18)

Substituting (18) into (12) yields

\[
w_s = \langle \phi \rangle \{ a \},
\] (19)

where

\[
\langle \phi \rangle = \langle P \rangle [C]^{-1}.
\] (20)

is the vector of the bending shape functions. In a similar way, shear deflection can be presented in the form

\[
w_s = \langle P \rangle \{ a \},
\] (21)

where according to (14)

\[
\langle P \rangle = -\frac{D}{S} \frac{\partial^2 \langle P \rangle}{\partial x^2} - \frac{D}{S} \frac{\partial^2 \langle P \rangle}{\partial y^2}.
\] (22)

Substituting (54) into (57) yields

\[
w_s = \langle \phi \rangle \{ a \},
\] (23)

where

\[
\langle \phi \rangle = \langle P \rangle [C]^{-1}.
\] (24)

is the vector of the shear shape functions.

Total deflection according to (8) reads

\[
w = \langle \phi \rangle \{ a \},
\] (25)

where

\[
\langle \phi \rangle = \langle \phi \rangle_\beta + \langle \phi \rangle_\gamma
\] (26)

is the vector of the total shape functions. Columns of the inverted matrix \([C]\) are vectors of coefficients \(a_i\) obtained for the unit value of particular nodal displacements

\[
[C]^{-1} = [\{ A \}_1, \{ A \}_2, \ldots, \{ A \}_n],
\] (27)

where

\[
\{ A \}_i = \{ a \}_i \{ a \}_i' \ldots \{ a \}_{n-1}.
\] (28)
Bending curvatures and warping are presented in the form

\[
\{\kappa\}_b = \begin{bmatrix}
\frac{\partial \lambda_{xx}}{\partial x} \\
\frac{\partial \lambda_{yy}}{\partial y} \\
\frac{\partial^2 \lambda_{xy}}{\partial x \partial y}
\end{bmatrix}
\]  

(29)

Substituting (19) with (20) into (29) yields

\[
\{\kappa\}_b = -[L]_b \{\sigma\},
\]

(30)

where

\[
[L]_b = [H]_b [C]^\dagger, \quad [H]_b = \begin{bmatrix}
\frac{\partial^2 \langle P \rangle_b}{\partial x^2} \\
\frac{\partial^2 \langle P \rangle_b}{\partial y^2} \\
2 \frac{\partial^2 \langle P \rangle_b}{\partial x \partial y}
\end{bmatrix}.
\]

(31)

Now it is possible to determine the bending stiffness matrix by employing a general formulation from the finite element method based on the variational principle

\[
[K]_b = \int [L]_b^T [D]_b [L]_b \, \text{d}A.
\]

(32)

where \([D]_b\) is the matrix of plate flexural rigidity. Furthermore, substituting (31) into (32) yields

\[
[K]_b = [C]^T [\mathcal{B}] [C]^\dagger.
\]

(33)

where symbolically \([C]^T = ([C]^\dagger)^T\) and

\[
[\mathcal{B}] = \int [H]_b^T [D]_b [H]_b \, \text{d}A.
\]

(34)

By taking (31) and (33) into account, (35) can be presented in the form

\[
[\mathcal{B}] = D ([I]_1 + v ([I]_2 + [I]_b) + [I]_d + 2(1 - v) [I]_b),
\]

(35)

where

\[
[I]_1 = \int \frac{\partial^2 \langle P \rangle_b}{\partial x^2} \, \text{d}A,
\]

(36)

\[
[I]_2 = \int \frac{\partial^2 \langle P \rangle_b}{\partial y^2} \, \text{d}A = [I]_1^T,
\]

(37)
Similarly, one can write for the shear strain vector

\[ \{\gamma\} = \begin{bmatrix} \frac{\partial \omega_x}{\partial x} \\ \frac{\partial \omega_y}{\partial y} \end{bmatrix} , \] (40)

\[ \{\gamma\} = [L]\{\delta\} , \] (41)

where

\[ [L] = [H]\{C\}^{-1} , \] (42)

\[ [H] = \begin{bmatrix} \frac{\partial (P)_x}{\partial x} \\ \frac{\partial (P)_y}{\partial y} \end{bmatrix} . \] (43)

Analogously to (32), the shear stiffness matrix based on the variational principle is presented in the form

\[ [K] = \int [L]^T [D] [L] \, dA , \] (44)

Where \([D] = S \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\). After some substitutions, according to [12], one can write

\[ [K] = [C]^T [S][C]^{-1} , \] (45)

\[ [S] = S \int [H]^T [H] \, dA , \] (46)

\[ [S] = S([I]_b + [I]_s) , \] (47)

\[ [I]_b = \int \frac{\partial^2 (P)_x}{\partial x^2} \, dA , \] (48)

\[ [I]_s = \int \frac{\partial^2 (P)_y}{\partial y^2} \, dA . \] (49)

Hence, the complete stiffness matrix is

\[ [K] = [K]_b + [K]_s = [C]^T ([B] + [S])[C]^{-1} . \] (50)

According to the general formulation of the mass matrix in the finite element method based on the variational principle one can write
\[ [M] = m \int_A \{ \phi \} \{ \phi \} dA, \quad (51) \]

where \( \{ \phi \} \) is the vector of the total shape functions (26). Taking (20) and (24) into account yields

\[ [M] = m \{ C \}^T [I_0] \{ C \}^{-1}, \quad (52) \]

Where \( [I_0] = \int_A \{ P \} \{ P \} dA \) and \( \{ P \} = \{ P \}_b + \{ P \}_s \), Eqs. (13) and (22).

4 NUMERICAL EXAMPLES – COMPARATIVE STUDY

Based on the above described theoretical background, an in-house code has been developed and applied to vibration analysis of square and rectangular plates, having different boundary conditions and different relative thicknesses. In all numerical examples values of Young's modulus is set at \( 2.1 \cdot 10^{11} \) N/m\(^2\), material density yields 7850 kg/m\(^3\), and the value of shear correction factor is set to 5/6. Results obtained by the developed code are denoted with PS (Present solution).

4.1 Simply supported plate, SSSS

Non-dimensional thickness parameters of simply supported square plate are presented in Table 1. They are compared with analytical solution, assumed mode method AMM [8] and NASTRAN values obtained by 2D and 3D FEM analysis [13]. Analytical results are rigorous since they also result from direct application of the Mindlin theory. The first 6 natural modes determined by 3D FEM NASTRAN analysis are shown in Figure 2 as relief map, where sagging and hogging modal areas are noticeable. FEM mesh used in 2D and 3D NASTRAN model is 8x8 and 8x8x4 elements, respectively. Boundary conditions for 3D FEM model are specified for nodes at the middle surface. In the considered case NASTRAN 2D and 3D frequency parameter values are very similar, and with PS values band the exact solution. The AMM values are very close to the analytical solution.

<table>
<thead>
<tr>
<th>MIN* ( m,n )</th>
<th>Analytical [4]</th>
<th>AMM</th>
<th>PS</th>
<th>NASTRAN 2D</th>
<th>NASTRAN 3D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,1</td>
<td>1.768</td>
<td>1.768</td>
<td>1.803</td>
<td>1.682</td>
<td>1.651</td>
</tr>
<tr>
<td>1,2; 2,1</td>
<td>3.866</td>
<td>3.869</td>
<td>4.024</td>
<td>3.791</td>
<td>3.717</td>
</tr>
<tr>
<td>2,2</td>
<td>5.588</td>
<td>5.591</td>
<td>5.827</td>
<td>5.249</td>
<td>5.142</td>
</tr>
<tr>
<td>1,3; 3,1</td>
<td>6.601</td>
<td>6.601</td>
<td>7.072</td>
<td>6.478</td>
<td>6.436</td>
</tr>
<tr>
<td>2,3; 3,2</td>
<td>7.974</td>
<td>7.975</td>
<td>8.466</td>
<td>7.336</td>
<td>7.273</td>
</tr>
</tbody>
</table>

*MIN - mode identification number

Table 1: Frequency parameter \( \lambda = \alpha^2 \sqrt{ph/D} / \pi^2 \) of square plate, SSSS, \( h/a=0.2 \)
4.2 Clamped square plate, CCCC

Vibration analysis of clamped plate is done by FEM using mesh of 8 x 8 elements. AMM, PS results and NASTRAN 2D and 3D results are listed in Table 2. In 3D FEM model all boundary nodes are fixed. The obtained results are compared with those from [5], which are obtained by the Rayleigh-Ritz method and due to high accuracy can be used for evaluation of the present solutions. AMM, as a variant of Rayleigh-Ritz method gives actually nearly the same results. As in the case of simply supported plate, PS and NASTRAN results band the referent values and discrepancies are of the same order of magnitude. The flexural natural modes of a clamped square plate determined by AMM and 2D FEM NASTRAN are shown in Fig. 3, and those determined by 3D FEM NASTRAN analysis are presented in Fig. 4.

Table 2: Frequency parameter $\lambda = \omega a^2 \sqrt{h / D / \pi^2}$ of square plate, CCCC, $h/a=0.2$

<table>
<thead>
<tr>
<th>Mode no.</th>
<th>Liew et al. [5]</th>
<th>AMM</th>
<th>PS</th>
<th>NASTRAN 2D</th>
<th>NASTRAN 3D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.687</td>
<td>2.688</td>
<td>2.724</td>
<td>2.657</td>
<td>2.758</td>
</tr>
<tr>
<td>2</td>
<td>4.691</td>
<td>4.691</td>
<td>4.842</td>
<td>4.607</td>
<td>4.787</td>
</tr>
<tr>
<td>3</td>
<td>4.691</td>
<td>4.691</td>
<td>4.842</td>
<td>4.607</td>
<td>4.787</td>
</tr>
<tr>
<td>5</td>
<td>7.177</td>
<td>7.177</td>
<td>7.640</td>
<td>6.942</td>
<td>7.240</td>
</tr>
<tr>
<td>6</td>
<td>7.276</td>
<td>7.276</td>
<td>7.697</td>
<td>7.013</td>
<td>7.357</td>
</tr>
</tbody>
</table>
4.3 Vibrations of rectangular plates

Further, rectangular plates with different combinations of boundary conditions and various relative thicknesses and length to width ratios are analysed. Values of non-dimensional frequencies parameters for SCCC, SCCS and FSSC boundary conditions are presented in Tables 3, 4 and 5, respectively. Comparisons with the results available in the relevant literature are also provided and similar comments as for square plates are valid. For illustration natural modes for SCCC and FSSC boundary conditions are shown in Figs. 5 and 6, respectively.

<table>
<thead>
<tr>
<th>$h/b$</th>
<th>METHOD</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>Xing &amp; Liu [17]</td>
<td>7.024</td>
<td>9.666</td>
<td>14.608</td>
<td>18.203</td>
</tr>
<tr>
<td></td>
<td>AMM</td>
<td>7.061</td>
<td>9.729</td>
<td>14.663</td>
<td>18.218</td>
</tr>
<tr>
<td></td>
<td>NASTRAN 2D</td>
<td>6.930</td>
<td>9.331</td>
<td>14.196</td>
<td>17.955</td>
</tr>
</tbody>
</table>

Table 3: Frequency parameter $\lambda = \omega^2 \sqrt{\rho h / D}$ of rectangular plate, SCCC, $a/b=0.6$

<table>
<thead>
<tr>
<th>$h/b$</th>
<th>METHOD</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>Xing &amp; Liu [17]</td>
<td>5.787</td>
<td>7.709</td>
<td>11.168</td>
<td>12.919</td>
</tr>
<tr>
<td></td>
<td>Liew et al. [5]</td>
<td>5.862</td>
<td>7.873</td>
<td>11.342</td>
<td>13.071</td>
</tr>
<tr>
<td></td>
<td>AMM</td>
<td>5.862</td>
<td>7.873</td>
<td>11.342</td>
<td>13.071</td>
</tr>
<tr>
<td></td>
<td>PS</td>
<td>5.851</td>
<td>7.823</td>
<td>11.480</td>
<td>13.252</td>
</tr>
<tr>
<td></td>
<td>NASTRAN 2D</td>
<td>5.775</td>
<td>7.534</td>
<td>10.706</td>
<td>12.825</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$h/b$</th>
<th>METHOD</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>Xing &amp; Liu [17]</td>
<td>4.154</td>
<td>5.472</td>
<td>7.725</td>
<td>8.293</td>
</tr>
<tr>
<td></td>
<td>Liew et al. [5]</td>
<td>4.282</td>
<td>5.687</td>
<td>7.900</td>
<td>8.373</td>
</tr>
<tr>
<td></td>
<td>AMM</td>
<td>4.282</td>
<td>5.687</td>
<td>7.900</td>
<td>8.373</td>
</tr>
<tr>
<td></td>
<td>PS</td>
<td>4.323</td>
<td>5.796</td>
<td>8.269</td>
<td>8.735</td>
</tr>
<tr>
<td></td>
<td>NASTRAN 2D</td>
<td>4.247</td>
<td>5.535</td>
<td>7.528</td>
<td>8.319</td>
</tr>
</tbody>
</table>

Table 4: Frequency parameter $\lambda = \omega^2 \sqrt{\rho h / D}$ of rectangular plate, SCCS, $a/b=0.6$
Figure 5: Mode shapes of rectangular plate, $a/b=0.6$, SCCC, AMM and NASTRAN 2D analysis

Figure 6: Mode shapes of rectangular plate, $a/b=0.4$, FSSC, AMM and NASTRAN 2D analysis

Table 5: Frequency parameter $\lambda = \omega h^2 \sqrt{\rho h / D}$ of rectangular plate, FSSC, $a/b=0.4$

<table>
<thead>
<tr>
<th>$h/b$</th>
<th>METHOD</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PS</td>
<td>9.896</td>
<td>11.172</td>
<td>14.017</td>
<td>18.746</td>
</tr>
<tr>
<td></td>
<td>NASTRAN 2D</td>
<td>9.823</td>
<td>10.945</td>
<td>13.363</td>
<td>17.404</td>
</tr>
<tr>
<td>0.1</td>
<td>Liew et al. [5]</td>
<td>7.941</td>
<td>8.970</td>
<td>11.135</td>
<td>14.462</td>
</tr>
<tr>
<td></td>
<td>AMM</td>
<td>8.055</td>
<td>9.068</td>
<td>11.206</td>
<td>14.508</td>
</tr>
<tr>
<td></td>
<td>PS</td>
<td>8.099</td>
<td>8.965</td>
<td>11.001</td>
<td>14.476</td>
</tr>
<tr>
<td></td>
<td>NASTRAN 2D</td>
<td>7.985</td>
<td>8.635</td>
<td>10.196</td>
<td>12.825</td>
</tr>
<tr>
<td>0.2</td>
<td>Liew et al. [5]</td>
<td>5.594</td>
<td>6.305</td>
<td>7.752</td>
<td>9.828</td>
</tr>
<tr>
<td></td>
<td>AMM</td>
<td>5.594</td>
<td>6.305</td>
<td>7.752</td>
<td>9.828</td>
</tr>
<tr>
<td></td>
<td>PS</td>
<td>5.765</td>
<td>6.386</td>
<td>7.885</td>
<td>10.325</td>
</tr>
<tr>
<td></td>
<td>NASTRAN 2D</td>
<td>5.669</td>
<td>6.126</td>
<td>7.277</td>
<td>8.993</td>
</tr>
</tbody>
</table>

4.4 Convergence study

The convergence of the proposed finite element formulation is demonstrated in the case of a
simply supported square plate. Natural frequencies are determined by the finite element model for three mesh densities, i.e. 6x6=36, 8x8=64 and 10x10=100 elements, and three values of a thickness ratio \( h/a \): 0.001, 0.1 and 0.2. The obtained frequency parameters are listed in Table 6 and compared with the exact analytical solution as well as with the NASTRAN and Abaqus results, [13,14]. In order to have better insight into the convergence, frequency parameter for the 1st, 4th and 7th modes are shown in Fig. 7. For thin plate (\( h/a=0.001 \)), the present solution converges to the exact value, faster than the NASTRAN and Abaqus results, which converge from the opposite sides. PS values for moderately thick plate (\( h/a=0.1 \)) are very close to the exact values for all three mesh densities, while the NASTRAN and Abaqus results converge to a lower value than the exact solution. The discrepancy is reduced for higher modes. In the case of thick plate (\( h/a=0.2 \)), the variation of the PS values, which are somewhat higher than the exact solution, is rather small. The NASTRAN and Abaqus results show the same tendency as in the previous case.

### Table 6: Convergence of frequency parameter \( \lambda = o\sqrt{\rho h/D} \) of square plate, SSSS

<table>
<thead>
<tr>
<th>h/a</th>
<th>METHOD</th>
<th>1/11/</th>
<th>2/12/</th>
<th>3/21/</th>
<th>4/22/</th>
<th>5/23/</th>
<th>6/24/</th>
<th>7/25/</th>
<th>8/26/</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>Senjanović et al. [4]</td>
<td>19.739</td>
<td>49.348</td>
<td>49.348</td>
<td>78.956</td>
<td>98.694</td>
<td>98.694</td>
<td>128.302</td>
<td>128.302</td>
</tr>
<tr>
<td></td>
<td>PS (6x6)</td>
<td>19.459</td>
<td>48.326</td>
<td>48.326</td>
<td>75.103</td>
<td>96.867</td>
<td>96.867</td>
<td>120.658</td>
<td>120.658</td>
</tr>
<tr>
<td></td>
<td>PS (8x8)</td>
<td>19.592</td>
<td>48.744</td>
<td>48.744</td>
<td>76.603</td>
<td>98.487</td>
<td>98.487</td>
<td>123.374</td>
<td>123.374</td>
</tr>
<tr>
<td></td>
<td>PS (10x10)</td>
<td>19.629</td>
<td>48.944</td>
<td>48.944</td>
<td>77.384</td>
<td>97.859</td>
<td>97.859</td>
<td>124.933</td>
<td>124.933</td>
</tr>
<tr>
<td></td>
<td>NASTRAN (6x6)</td>
<td>19.150</td>
<td>47.343</td>
<td>47.343</td>
<td>71.762</td>
<td>95.014</td>
<td>95.014</td>
<td>115.491</td>
<td>115.491</td>
</tr>
<tr>
<td></td>
<td>NASTRAN (8x8)</td>
<td>19.377</td>
<td>48.115</td>
<td>48.115</td>
<td>74.420</td>
<td>96.381</td>
<td>96.381</td>
<td>119.255</td>
<td>119.255</td>
</tr>
<tr>
<td></td>
<td>NASTRAN (10x10)</td>
<td>19.488</td>
<td>48.497</td>
<td>48.497</td>
<td>75.803</td>
<td>97.063</td>
<td>97.063</td>
<td>122.044</td>
<td>122.044</td>
</tr>
<tr>
<td></td>
<td>Abaqus (6x6)</td>
<td>20.205</td>
<td>54.743</td>
<td>54.743</td>
<td>86.623</td>
<td>132.223</td>
<td>132.223</td>
<td>158.355</td>
<td>158.355</td>
</tr>
<tr>
<td></td>
<td>Abaqus (8x8)</td>
<td>20.000</td>
<td>52.251</td>
<td>52.251</td>
<td>83.203</td>
<td>115.389</td>
<td>115.389</td>
<td>143.839</td>
<td>143.839</td>
</tr>
<tr>
<td></td>
<td>Abaqus (10x10)</td>
<td>19.906</td>
<td>51.169</td>
<td>51.169</td>
<td>81.652</td>
<td>124.895</td>
<td>124.895</td>
<td>137.892</td>
<td>137.892</td>
</tr>
<tr>
<td>0.1</td>
<td>Senjanović et al. [4]</td>
<td>19.065</td>
<td>45.483</td>
<td>45.483</td>
<td>69.794</td>
<td>85.038</td>
<td>85.038</td>
<td>106.684</td>
<td>106.684</td>
</tr>
<tr>
<td></td>
<td>PS (6x6)</td>
<td>18.959</td>
<td>45.581</td>
<td>45.581</td>
<td>69.055</td>
<td>87.636</td>
<td>87.636</td>
<td>107.334</td>
<td>107.334</td>
</tr>
<tr>
<td></td>
<td>PS (8x8)</td>
<td>19.063</td>
<td>45.821</td>
<td>45.821</td>
<td>69.908</td>
<td>87.320</td>
<td>87.320</td>
<td>108.019</td>
<td>108.019</td>
</tr>
<tr>
<td></td>
<td>PS (10x10)</td>
<td>19.117</td>
<td>45.949</td>
<td>45.949</td>
<td>70.374</td>
<td>87.230</td>
<td>87.230</td>
<td>108.520</td>
<td>108.520</td>
</tr>
<tr>
<td></td>
<td>NASTRAN (6x6)</td>
<td>18.022</td>
<td>42.985</td>
<td>42.985</td>
<td>60.972</td>
<td>80.069</td>
<td>80.069</td>
<td>89.787</td>
<td>89.787</td>
</tr>
<tr>
<td></td>
<td>NASTRAN (8x8)</td>
<td>18.243</td>
<td>43.927</td>
<td>43.927</td>
<td>64.251</td>
<td>82.692</td>
<td>82.692</td>
<td>96.541</td>
<td>96.541</td>
</tr>
<tr>
<td></td>
<td>NASTRAN (10x10)</td>
<td>18.339</td>
<td>44.380</td>
<td>44.380</td>
<td>65.918</td>
<td>83.926</td>
<td>83.926</td>
<td>100.115</td>
<td>100.115</td>
</tr>
<tr>
<td></td>
<td>Abaqus (6x6)</td>
<td>18.906</td>
<td>49.078</td>
<td>49.078</td>
<td>73.469</td>
<td>106.435</td>
<td>106.435</td>
<td>122.766</td>
<td>122.766</td>
</tr>
<tr>
<td></td>
<td>Abaqus (8x8)</td>
<td>18.663</td>
<td>46.971</td>
<td>46.971</td>
<td>70.836</td>
<td>95.661</td>
<td>95.661</td>
<td>114.072</td>
<td>114.072</td>
</tr>
<tr>
<td></td>
<td>Abaqus (10x10)</td>
<td>18.544</td>
<td>46.037</td>
<td>46.037</td>
<td>69.609</td>
<td>91.190</td>
<td>91.190</td>
<td>110.279</td>
<td>110.279</td>
</tr>
<tr>
<td>0.2</td>
<td>Senjanović et al. [4]</td>
<td>17.449</td>
<td>38.152</td>
<td>38.152</td>
<td>55.150</td>
<td>65.145</td>
<td>65.145</td>
<td>78.697</td>
<td>78.697</td>
</tr>
<tr>
<td></td>
<td>PS (6x6)</td>
<td>17.662</td>
<td>39.476</td>
<td>39.476</td>
<td>56.997</td>
<td>70.314</td>
<td>70.314</td>
<td>83.630</td>
<td>83.630</td>
</tr>
<tr>
<td></td>
<td>PS (8x8)</td>
<td>17.733</td>
<td>39.454</td>
<td>39.454</td>
<td>57.016</td>
<td>69.123</td>
<td>69.123</td>
<td>82.667</td>
<td>82.667</td>
</tr>
<tr>
<td></td>
<td>PS (10x10)</td>
<td>17.767</td>
<td>39.451</td>
<td>39.451</td>
<td>57.081</td>
<td>68.584</td>
<td>68.584</td>
<td>82.280</td>
<td>82.280</td>
</tr>
<tr>
<td></td>
<td>NASTRAN (6x6)</td>
<td>16.348</td>
<td>36.264</td>
<td>36.264</td>
<td>48.588</td>
<td>60.567</td>
<td>60.567</td>
<td>65.908</td>
<td>65.908</td>
</tr>
<tr>
<td></td>
<td>NASTRAN (8x8)</td>
<td>16.353</td>
<td>37.148</td>
<td>37.148</td>
<td>51.336</td>
<td>63.253</td>
<td>63.253</td>
<td>71.578</td>
<td>71.578</td>
</tr>
<tr>
<td></td>
<td>NASTRAN (10x10)</td>
<td>16.617</td>
<td>37.562</td>
<td>37.562</td>
<td>52.679</td>
<td>64.503</td>
<td>64.503</td>
<td>74.427</td>
<td>74.427</td>
</tr>
<tr>
<td></td>
<td>Abaqus (6x6)</td>
<td>16.796</td>
<td>40.014</td>
<td>40.014</td>
<td>56.260</td>
<td>76.247</td>
<td>76.247</td>
<td>85.787</td>
<td>85.787</td>
</tr>
<tr>
<td></td>
<td>Abaqus (8x8)</td>
<td>16.610</td>
<td>38.620</td>
<td>38.620</td>
<td>54.749</td>
<td>70.632</td>
<td>70.632</td>
<td>81.596</td>
<td>81.596</td>
</tr>
<tr>
<td></td>
<td>Abaqus (10x10)</td>
<td>16.523</td>
<td>37.997</td>
<td>37.997</td>
<td>54.048</td>
<td>68.169</td>
<td>68.169</td>
<td>79.675</td>
<td>79.675</td>
</tr>
</tbody>
</table>

*\( m, n \) - mode identification number, \( m \) and \( n \) number of half waves in \( x \) and \( y \) direction
Figure 7: Convergence of frequency parameter $\lambda = \omega^2 \sqrt{\rho h / D}$ of simply supported square plate

5 CONCLUDING REMARKS

Newly developed shear locking-free finite element for vibration analysis of moderately thick plate is presented and validated with in extensive numerical calculations. The results are compared to recently published analytical solutions, assumed mode method, Rayleigh-Ritz solutions and some general commercial FE software results. Decomposition of total deflection into bending and shear components and reduction of three differential equations into one single
with bending deflection as a potential function, provides clear physical insight into the problem. Moreover, when the FE is derived using the proposed procedure based on the advanced theory of thick plate vibrations, shear locking phenomenon doesn’t appear. Therefore usual procedure for shear locking elimination are not required in this case.

ACKNOWLEDGEMENTS

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIP) through GCRC-SOP (Grant No. 2011-0030013).

REFERENCES

DYNAMIC ANALYSIS BY A NEW FAMILY OF TIME-MARCHING PROCEDURES WITH ADAPTIVE TIME INTEGRATION PARAMETERS

DELFIM SOARES

Structural Engineering Department, Federal University of Juiz de Fora
Cidade Universitária, CEP 36036-330, Juiz de Fora, MG, Brazil
E-mail: delfim.soares@ufjf.edu.br

Key words: Time-Marching; Time Integration Methods; Dynamics; Accuracy; Stability.

Abstract. In this work, a new family of time-marching procedures for dynamics is discussed. The new methodology is based only on displacements/velocities relations, requiring no computation of accelerations. Expressions for the time integration parameters of the method, which adapt according to the properties of the model, are suggested, rendering a very effective technique. In fact, the new procedure provides much more accurate results than standard time-marching techniques, being quite competitive. The main characteristics of the new family are: (i) it is a very simple technique and it can be easily implemented into any existing code; (ii) some well-known and widely used methods can be reproduced by the technique; (iii) it is truly self-starting, requiring no initial procedures at all; (iv) it is second-order accurate; (v) it has controllable period elongation properties; (vi) it has controllable algorithmic dissipation; (vii) it is unconditionally stable when implicit analysis are focused; (viii) it allows explicit analysis to be carried out; (ix) it is efficient and no more than one set of implicit equations needs to be solved at each step. Along the paper, numerical results are presented, illustrating the good performance and high potentialities of the new method.

1 INTRODUCTION

The literature reports many classical explicit [1-3] and implicit [4-6] algorithms for time-marching analysis (for a comprehensive review, see [7]). Explicit procedures are usually preferable because of their lower computational effort, being the restrictions due to stability conditions their main disadvantage. Implicit approaches, on the other hand, can be settled unconditionally stable, being characterized, however, by higher computational costs. Many procedures can be employed to improve stability and accuracy of time-integration algorithms, such as subcycling techniques, high-order accurate schemes, automatic time step control etc. As a matter of fact, a lot of research is continuously realized on this field and several time-marching techniques are available nowadays for dynamic analysis [8-15].

In this work, a new family of time-marching procedures is presented. This new family is based on a simple weighted residual form of the equation of motion and it deals only with displacements and velocities relations, being the computation of accelerations not required by
the technique. Thus, the methodology is very simple and truly self-starting. As it is described along the text, some classical widely used time integration methods can be reproduced by the new approach, such as the trapezoidal rule. However, special features of the new technique may be explored, and enhanced procedures enabled. In the new technique, period elongation errors can be controlled, allowing creating a very accurate methodology. Additionally, it is second-order accurate, it has controllable algorithmic dissipation and it is quite efficient, requiring no more than one set of implicit equations to be solved at each time-step. Explicit approaches can also be reproduced by the technique, eliminating the solution of any system of equations per time-step.

2 GOVERNING EQUATIONS AND TIME INTEGRATION STRATEGY

The governing system of equations describing a linear dynamic model is given by [16]:

\[ \mathbf{M}\ddot{\mathbf{U}}(t) + \mathbf{C}\dot{\mathbf{U}}(t) + \mathbf{K}\mathbf{U}(t) = \mathbf{F}(t) \]  

(1)

where \( \mathbf{M}, \mathbf{C} \) and \( \mathbf{K} \) are mass, damping and stiffness matrices, respectively, \( \mathbf{F}(t) \) stands for the force vector and \( \mathbf{U}(t), \dot{\mathbf{U}}(t) \) and \( \ddot{\mathbf{U}}(t) \) are displacement, velocity and acceleration vectors, respectively. The initial conditions of the model are given by \( \mathbf{U}^0 = \mathbf{U}(0) \) and \( \dot{\mathbf{U}}^0 = \dot{\mathbf{U}}(0) \), where \( \mathbf{U}^0 \) and \( \dot{\mathbf{U}}^0 \) stand for initial displacement and velocity vectors, respectively.

By time integrating the equation of motion (1), considering a time-step \( \Delta t \), one may write:

\[ \int_{t_n}^{t_{n+1}} \mathbf{M}\ddot{\mathbf{U}}(\tau)\,d\tau + \int_{t_n}^{t_{n+1}} \mathbf{C}\dot{\mathbf{U}}(\tau)\,d\tau + \int_{t_n}^{t_{n+1}} \mathbf{K}\mathbf{U}(\tau)\,d\tau = \int_{t_n}^{t_{n+1}} \mathbf{F}(\tau)\,d\tau \]  

(2)

which may be viewed as a simple weighted residual form of the equation of motion.

The integrals in the l.h.s. of equation (2) may be approximated by:

\[ \mathcal{J}_{U}^{n+\frac{1}{2}} = \int_{t_n}^{t_{n+\frac{1}{2}}} \ddot{\mathbf{U}}(\tau)\,d\tau \approx \ddot{\mathbf{U}}^{n+1} - \ddot{\mathbf{U}}^n \]  

(3a)

\[ \mathcal{J}_{\dot{U}}^{n+\frac{1}{2}} = \int_{t_n}^{t_{n+\frac{1}{2}}} \dot{\mathbf{U}}(\tau)\,d\tau \approx \dot{\mathbf{U}}^{n+1} - \dot{\mathbf{U}}^n \]  

(3b)

\[ \mathcal{J}_{U}^{n+1} = \int_{t_n}^{t_{n+1}} \mathbf{U}(\tau)\,d\tau \approx \Delta t\,\mathbf{U}^n + \frac{1}{2} \alpha_2 \Delta t^2 \mathbf{U}^n + \frac{1}{2} \alpha_1 \Delta t^2 \ddot{\mathbf{U}}^{n+1} \]  

(3c)

where \( \alpha_1 \) and \( \alpha_2 \) are parameters of the new method. The displacement \( \mathbf{U}^{n+1} \) can also be defined by the following simple finite difference expression:

\[ \mathbf{U}^{n+1} = \mathbf{U}^n + \delta_2 \Delta t \ddot{\mathbf{U}}^n + \delta_1 \Delta t \dot{\mathbf{U}}^{n+1} \]  

(4)

where \( \delta_1 \) and \( \delta_2 \) are additional parameters of the method.

Taking into account approximations (3) and (4), equation (2) may be rewritten as the following recursive relation:
\[(M + \delta_1 \Delta t C + \frac{i}{2} \alpha_1 \Delta t^2 K) \mathbf{U}_{n+1} = \mathbf{Z}_{n+\frac{1}{2}} + \mathbf{M} \dot{\mathbf{U}}_{n} - \delta_2 \Delta t C \dot{\mathbf{U}}_{n} - \mathbf{K}(\Delta t \mathbf{U}_{n} + \frac{i}{2} \alpha_2 \Delta t^2 \ddot{\mathbf{U}}_{n})\]  

where \(\mathbf{Z}_{n+\frac{1}{2}}\) stands for the integral in the r.h.s. of equation (2).

Equation (5) enables to compute the velocities \(\dot{\mathbf{U}}_{n+1}\), and equation (4) may then be used to compute the displacements at the current time step. It is important to highlight that the method described by equations (4) and (5) is based only on velocities and displacements, being no computation of accelerations required. Thus, the first positive feature of the method is that it is truly self-starting, eliminating any kind of cumbersome initial procedures, such as the computation of initial accelerations (which usually requires an extra system of equations to be dealt with) and/or the computation of multistep initial values.

In this work, two major groups of time-marching procedures are focused, namely: (i) explicit conditionally stable techniques and (ii) implicit unconditionally stable techniques. As it is very well known, both these groups have advantages and disadvantages and they are more commonly used according to the application in focus. Explicit conditionally stable techniques, for instance, do not require any system of equations to be solved along the time marching procedure, being very efficient; however, they are limited by a critical time-step, above which instability occurs. Implicit unconditionally stable techniques, on the other hand, may consider higher time-steps, being limited just by accuracy criteria; however, they require a system of equations to be solved at each time-step, which can be very computationally demanding in some cases.

Taking into account these two groups, the following expressions are here suggested for the time integration parameters of the method:

\[\delta_1 = \frac{1}{2} + d \omega \Delta t; \quad \alpha_1 = 0; \quad \delta_2 = 1 - \delta_1; \quad \alpha_2 = 2 \delta_2\]  

\[\delta_1 = \frac{1}{2} + d \omega \Delta t; \quad \alpha_1 = \frac{1}{2} + \frac{3}{2} \tanh(a \omega \Delta t); \quad \delta_2 = 1 - \delta_1; \quad \alpha_2 = 2(2 \alpha_1)^{1/2} - \alpha_1 - 2 \delta_1\]  

where equations (6) are related to explicit analyses, whereas equations (7) are related to implicit analyses. In these equations, \(d\) stands as a parameter which controls the accuracy of the technique (it controls period elongation errors) and \(a\) stands as a parameter which controls the numerical dissipation of the technique. \(\omega\) represents an approximation for the lowest natural frequency of the model; thus, as one can observe, the suggested time integration procedure will adapt itself according to the properties of the model, once the time integration parameters are function of the main eigenvalue of the problem.

In equations (6), if \(d = 0\) is considered, the basic characteristics of the central difference method are reproduced. In equations (7), if \(d = 0\) and \(a = 0\) are considered, the trapezoidal rule (or constant average acceleration method) is reproduced. Following expressions (6) or (7), the method becomes second order accurate, no matter the adopted values for \(d\), \(a\) or \(\omega\). The methodology described by equations (6) is conditionally stable, with its critical time step being the same of the central difference method. The methodology described by equations (7) is unconditionally stable.

In Fig.1, period elongation, amplitude decay and amplitude factor errors are depicted, considering expressions (6) and (7), illustrating the good accuracy of the technique. In Fig.2, spectral radii are depicted, considering expressions (7) (following expressions (6), the spectral
behavior of the central difference method is reproduced). As one can observe in the figures, the new technique is more accurate than the central difference method (for the explicit case, \( d = -1/4 \) is suggested) or the trapezoidal rule (for the implicit case, \( d = 1/2 \) is suggested). Numerical damping is also very properly introduced into the technique, as described in Fig.2. In this case, high spurious modes may be properly extinguished, since the proposed technique is \( L \)-stable (i.e., \( \rho_c = 0 \)). Additionally, as previously remarked, second-order accuracy is maintained when numerical dissipation is introduced into the analysis (i.e., when \( a > 0 \)).

In Fig.3, error results versus time discretization are presented considering a single degree of freedom model submitted to a unitary displacement initial condition, further illustrating the superior accuracy of the technique once proper values for \( d \) are selected.

---

**Figure 1**: Period elongation, amplitude decay and amplitude factor for the suggested technique: first column − explicit analysis; second column − implicit analysis with \( a = 0 \); third column − implicit analysis with \( a = 0.01 \).
For further details about the method and a deeper mathematical analysis of its properties, the work of Soares [15] is referred. In the next section, some numerical applications are considered taking into account multi degree of freedom models, further illustrating the good performance of the technique.

3 NUMERICAL APPLICATIONS

In this section, plane elastic models, discretized by the finite element method, are considered. First, a clamped rod, submitted to axial forces, is analyzed. A sketch of the model is depicted in Fig.4a. The adopted finite element discretization is also described in the figure. In this case, 160 linear triangular elements are employed to spatially discretize the rod. The physical properties of the model are: $E = 100.0$ N/m$^2$ (Young modulus); $\nu = 0.0$ (Poisson
coefficient); and \( \rho = 1.0 \text{ kg/m}^3 \) (mass density). The geometry of the model is defined by \( L = 1.0 \) m. Considering that the axial forces are suddenly applied and kept constant along time (Heaviside time behaviour), analytical answers for the horizontal displacement of the model are available [15]. Thus, the computed results can be compared to analytical answers, and errors can be evaluated.

\[
\begin{align*}
\text{Figure 4: Sketch of the MDOF models: (a) homogeneous rod; (b) heterogeneous beam.}
\end{align*}
\]

\[
\begin{align*}
\text{Figure 5: Error vs. time discretization for the rod: (a) } d = 0.3; \text{ (b) } d = 0.5; \text{ (c) } d = 0.7.
\end{align*}
\]

In Fig. 5, error results are depicted, taking into account the middle point of the rod and a total time of analysis of 20s. Results are depicted considering the trapezoidal rule (grey lines) and expressions (7) (black lines). To compute \( \omega \) in equations (7), the simple expression \( \omega \approx [\min(K_{ii}/M_{ii})]^{1/2} \) is considered (i.e., \( \omega \) is computed as the square root of the minimal value of the diagonal terms of the stiffness matrix divided by the diagonal terms of the mass matrix). This is, of course, an extremely poor approximation for the lower natural frequency of the model; however, this poor approximation is considered here to highlight the good performance of the method, even considering an eventually poor estimation for \( \omega \). As one can observe in Fig. 5, very accurate results are obtained by the proposed technique.

It is important to highlight that the oscillatory behaviour of the curves in Fig. 5 is expected. In this case, the convergence of the time-marching procedure is towards the finite element spatial discretization solution, and not towards the analytical space-time solution (one should keep in mind that the finite element spatial discretization also introduces errors). Thus, oscillations usually arise in these error curves when errors are computed considering time integration algorithms in conjunction with finite element spatial discretizations, and the reference response is the analytical answer of the model.
As it is described in Fig.5, better results are obtained if some numerical damping is introduced into the analysis. In this case, not only high frequency modes can be dissipated, but also amplitude factor errors can be compensated.

In Fig.4b, a sketch of a heterogeneous clamped beam is depicted. In the first 30% of the model, the Young modulus of the beam is 2000 N/m²; in the next 30%, it is 500 N/m², and in the remaining 40%, it is 100 N/m². The mass density and the Poisson coefficient are 1.0 kg/m² and 0.25, respectively, for the entire model. The same finite element mesh of the rod is employed to discretize the beam.

No analytical answers are available for this heterogeneous model, and the results obtained by the central difference method with $\Delta t = 10^{-5}$ s are considered as a reference response. In Fig.6, error results are depicted, taking into account the middle point of the beam and a total time of analysis of 20 s. In Figs.6a and b, explicit (expressions (6)) and implicit (expressions (7)) analyses are focused, respectively.

As one can observe, once again very good accuracy is obtained by the proposed technique. For explicit analyses, since very small time-steps must be considered in order to ensure stability, lower differences between the computed errors are observed. In implicit analyses, on the other hand, these differences may be amplified, once larger time-steps can be employed. As it is illustrated in Fig.6, the new technique is more accurate than the central difference method (explicit analyses) and the trapezoidal rule (implicit analyses). In the Newmark method, for $\delta > \frac{1}{2}$, numerical damping is introduced and the method becomes first-order accurate. In the new technique, by choosing $\alpha > 0$, numerical damping is introduced, maintaining second-order accuracy. Period elongation errors are also reduced by choosing $d > 0$ (Fig.6b), improving the accuracy of the analysis and keeping second-order accuracy.

As it can be observed, the new methodology is very effective, allowing computing very accurate results with low computational costs.

**Figure 6**: Error vs. time discretization for the beam: (a) explicit analysis; (b) implicit analysis.
4 CONCLUSIONS

In this work, a new family of time-marching procedures for hyperbolic models is presented. Basically, the main advantages of the proposed new family of time-marching procedures are: (i) it is a very simple technique and it can be easily implemented into any existing code; (ii) some well-known and widely used methods can be reproduced by the technique; (iii) it is truly self-starting, requiring no initial procedures at all; (iv) it is second-order accurate; (v) it has controllable period elongation properties; (vi) it has controllable algorithmic dissipation; (vii) it is unconditionally stable when implicit analysis are focused; (viii) it allows explicit analysis to be carried out; (ix) it is efficient and no more than one set of implicit equations needs to be solved at each step.

As it was illustrated in section 3, the accuracy of the new technique is remarkable. The method allows controlling period elongation errors, providing very accurate results. One of the main positive features of the new technique is that it may control stability (\( \alpha \) parameters) and accuracy (\( \delta \) parameters) aspects independently and, as a consequence, these aspects can be better optimized. Thus, a superior performance can be achieved, as it is illustrated along this work. As a matter of fact, by comparing the curves in Fig.1 with those of standard time-marching procedures, the superior performance of the proposed technique becomes very clear.

ACKNOWLEDGEMENTS

The financial support by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) and FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais) is greatly acknowledged.

REFERENCES


DIRECT ASSESSMENT OF INTERLAMINAR STRESSES IN COMPOSITE MULTILAYERED PLATES USING A LAYER-WISE MIXED FINITE ELEMENT MODEL

ROBERTO S. CARNICER¹, BRAIAN A. DESIA¹ AND RODOLFO A. SCHWARZ¹

¹Faculty of Engineering, University of Belgrano
Villanueva 1324, Autonomous City of Buenos Aires, Argentina
rcarnicer@freyreyasoc.com.ar, b.desia@hotmail.com.ar, rodolfo.schwarz@comunidad.ub.edu.ar

Key words: Interlaminar stresses, Multilayered plates, Mixed finite element, Lagrange multipliers.

Abstract. An accurate determination of interlaminar transversal stresses in composite multilayered plates, especially near free-edge, is of great importance in the study of inter-ply damage modes, mainly in the initiation and growth of delamination. In this paper, interlaminar stresses are determined by layer-wise mixed finite element model. Each layer is analyzed as an isolated one where the displacement continuity is ensured by means Lagrange Multipliers (which represent the statics variables). This procedure allows us to work with any single plate model, obtaining the interlaminar stresses directly without loss of precision. The FSDT with transverse normal strain effects included is assumed in each layer, but Lagrange Polynomials are used to describe the kinematic instead of Taylor’s polynomial functions of the thickness coordinates, as is common. This expansion allows us to pose the interlaminar displacements compatibility simpler than the second one. The in-plane domain of the plate is discretized by four-node quadrilateral elements, both to the field of displacement and to the Lagrange multipliers. The mixed interpolation of tensorial components technique is applied to avoid the shear-locking in the finite element model. Several examples were carried out and the results have been satisfactorily compared with those available in the literature.

1 INTRODUCTION

Advanced composite materials are widely used in many branches of the engineering, such as civil, naval, aeronautic and aerospace industries, because of their better properties compared to conventional materials. Among the different types of composite materials, laminates are the most popular due to variety of structural applications. Composite laminates are formed by stacking of different layers. Since the laminas are linked through their faces, transverse shear stresses are developed on the interface between two layers. Although these stresses are negligible in the regions away from the laminate boundaries, generally transverse stress (σxz, σyz, σzz) concentrations are exhibited near material and geometric discontinuities (the so-called free-edge effect) that can lead to damage in the form of delamination, matrix cracking, and adhesive joint separation, causing loss of stiffness and a complex load redistribution [1]. The study of this problem arises around 1970 when many laminate test coupons loaded in axial tension exhibited unusual failure modes, precipitated near the free
edges. By this time, Pipes and Pagano [2, 3] developed an elastic solution, based on finite difference technique, for finite width composite laminates under uniform axial strain. Their observations have led to the understanding of the interlaminar stresses in the region near the free edge of a laminate. They showed strong evidence of a singularity in the interlaminar $\sigma_{xz}$ and $\sigma_{zz}$ at the intersection of the interface and free-edge for angle-ply laminates $[\pm45^\circ]$. These high stresses in the neighborhood of the free edge may be expected to cause delamination of the laminate, in particular under fatigue loadings. At present, a three-dimensional elasticity exact solution to evaluate the nature of stress concentration near edge boundary in composite laminates is unavailable yet. A complete review of literature is given by Kant and Swaminathan [4].

Thus, a challenging task in composite plates/shells modelling is obtain an accurate estimation of interlaminar transverse stresses \textit{a priori} (i.e., without requiring any post-processing procedure, such as integration of the three-dimensional equilibrium equations, that leads to loss of precision). This paper presents a numerical solution based on a layered model for the direct determination of interlaminar stresses via finite elements with mixed variables (kinematics and statics).

The paper has been organized as follows: Section 2 provides the mathematical formulation of the problem, including the kinematics and the strong form of the equilibrium equations. Section 3 develops the finite element model from the weak form of the governing equations. The methodology to solve just the static variables is shown, based on a strategy to eliminate the kinematic unknowns. Several examples are solved and exposed in Section 4. Finally, Section 6 contains the conclusions.

2 PROBLEM FORMULATION

Throughout, Greek and Capital Latin indexes take on the values 1,2,3 and 1,2, respectively. Also, the repeated indexes are summed over their ranges, i.e., Einstein Notation Convention is adopted.

Infinitesimal strain and small displacement are assumed. Due to this, no distinction is made between the material coordinates and spatial coordinates, between the Green-Lagrange strain tensor and infinitesimal strain tensor, and between the Second Piola-Kirchhoff stress tensor and the Cauchy stress tensor.

2.1 Kinematics

Consider a plate of total thickness $h$ composed of $n$ orthotropic layers which are perfectly bonded together. A Cartesian reference system is defined in the undeformed midplane $\Omega$ of the plate, where $(x, y)$ are the in-plane coordinates (see Figure 1). The $z$-axis is taken positive downward from the midplane. The generic $i^{th}$ layer is located between the points $z_{\text{top}}=z_i^+$ and $z_{\text{bot}}=z_i^-$ in the thickness direction. For the purpose of modelling, all layers are assumed here to have the same in-plane dimensions (i.e., it is assumed that $\Omega_i=\Omega$) but the thickness $h_i$ of each might differ. The middle surface $\Omega_i$ of each layer is bounded by $\Gamma_i$ ($\Gamma_i^\sigma$ and $\Gamma_i^\rho$ refer those parts of $\Gamma_i$ on which the geometrical and mechanical boundary conditions are prescribed, respectively). The integer $i$, which is extensively used as both subscripts or superscripts, denotes the layer number that starts from the plate top. Symbols that are not affected by the index $i$ refer to the whole plate.
Using Carrera’s Unified Formulation (CUF) [5], the displacement field \( \{u_{\alpha}, \alpha=1,2,3\} \) for any point \( \{x,y,z\} \) of the \( i^{th} \) layer can be expressed in a compact notation as follows,

\[
\begin{align*}
    u_{\alpha}^{(i)}(x,y,z) &= F_k(z) u_{\alpha}^{(i)}(x,y) \\
    k &= 1, \ldots, N^* 
\end{align*}
\]

where \( F_k(z) \) is the set of thickness functions and \( N^* \geq 2 \) is the order of the expansion, which can both be arbitrarily chosen. In the present work, a linear thickness expansion \( (N^*=2) \) based on Lagrange polynomial is used. Furthermore, the same functions are used for all components of the unknown field variables. Thus, in explicit way,

\[
\begin{align*}
    u_1^{(i)} &= F_t u_1^{(i)} + F_b u_1^{(i)} \\
    u_2^{(i)} &= F_t u_2^{(i)} + F_b u_2^{(i)} \\
    u_3^{(i)} &= F_t u_3^{(i)} + F_b u_3^{(i)}
\end{align*}
\]

in which the thickness function \( F_k(z) \) are,

\[
\begin{align*}
    F_t(z) &= \frac{1+\zeta(z)}{2} \\
    F_b(z) &= \frac{1-\zeta(z)}{2}
\end{align*}
\]

The dimensionless thickness coordinate \( \zeta(z) \in [-1;+1] \) is introduced at layer level by the following relation,

\[
\zeta_i(z) = -\frac{2}{h_i} z + \frac{z_i + z_{i+1}}{h_i}
\]

The chosen functions have the following interesting properties,

\[
\zeta_i = \begin{cases} +1 : F_t = 1; F_b = 0 \\
-1 : F_t = 0; F_b = 1 
\end{cases}
\]

Notice that the top and bottom values (denoted by \( t \) and \( b \), respectively) are used as unknowns, instead of variables referred to the midplane as usual in Taylor expansion. This
representation is more convenient for the purpose of implementing the interlaminar continuity between two adjacent layers (i.e., local displacements \( \{ u_\alpha, \alpha = 1,2,3 \} \) at the bottom of the layer \( i \) and the top of the layer \( i+1 \), have to be continuous). In a Taylor polynomial thickness expansion for a displacement variable described in a layer-wise sense is not convenient since many degrees of freedom are involved in the writing of the interlaminar compatibility condition. Based on (1), the strong form of such a compatibility condition is,

\[
 u^{(i+1)}_{\alpha t} - u^{(i)}_{\alpha b} = 0 
\]

(6)

The Cartesian components of the infinitesimal strain tensor at level layer are,

\[
 \varepsilon_{,k}^{(i)} = \frac{1}{2} F_k \left( u^{(i)}_{A k} + u^{(i)}_{B k} \right) 
\]

(7a)

\[
 \varepsilon_{33}^{(i)} = \frac{1}{2} \left( F_k u^{(i)}_{3 k A} + F_k u^{(i)}_{3 k A} \right) 
\]

(7b)

\[
 \varepsilon^{(i)}_{33} = F_k u^{(i)}_{3 k} 
\]

(7c)

where the comma indicates the spatial partial differentiation. Note that the transverse normal strain \( \varepsilon_{33} \) is included in the kinematics. Models without transverse normal strain are no capable of accurately determining interlaminar stresses near geometric and material discontinuities such as holes or cut-outs, traction free edges, delamination fronts and matrix crack fronts, or in localized regions of intense loading. In modelling these localized effects, inclusion of the transverse normal strain is important for two reasons. First of all, the transverse normal stress is usually a significant, if not dominant, stress in these regions (perhaps even singular) (see [3]). Secondly, layer-wise models that neglect transverse normal strain do not satisfy traction-free boundary conditions for transverse shear stresses at the laminate edges [1].

2.2 Governing equations

Multilayered system is obtained as a superposition of \( n \) ply, where the interlaminar continuity requirements (6) are mathematically ensured by means Lagrange multipliers \( \{ s^{(i)}_{\alpha}, \alpha = 1,2,3 \} \) which represents surfaces forces distributed along of interfaces between two layers. Thus, the potential energy functional \( \Pi_{mP1} \) is expressed as,

\[
 \Pi_{mP1} = \sum_{i=1}^{n} \left\{ \int_{V_i} \frac{1}{2} \sigma_{ab}^{(i)} \varepsilon_{ab}^{(i)} \, dV \right\} + \\
 + \sum_{j=2}^{n} \left\{ \int_{S_{j+1}} \left( u^{(j)}_{\alpha, \text{top}} - u^{(j-1)}_{\alpha, \text{bot}} \right) s^{(i)}_{\alpha} \, dS \right\} - W^{\text{ext}} \quad k = t, b
\]

(8)

where \( \sigma_{ab} \) are the Cartesian components of the Cauchy stress tensor and \( W^{\text{ext}} \) is the work done by the external layer-forces \( \{ p^{(i)}_{0k}, \alpha = 1,2,3, k = t, b \} \). The \( i \)-layer volume is denoted as \( V_i \), while \( S_{j+1} \) refers to the interface between the layers \( j \) and \( j+1 \). The other quantities have already been defined previously. Note that the first term in the right-hand side of (8) represents the internal work in each layer, and the second one variationally enforces the subsidiary conditions (6). The Lagrange multipliers introduced have the following physical interpretation [6] (see Figure 2),

\[
 s^{(i)}_{A} = \sigma^{(i)}_{33} 
\]

(9a)

\[
 s^{(i)}_{3} = \sigma^{(i)}_{33} 
\]

(9b)
Indeed, the interface conditions (6) could be imposed in a procedure similar to standard assembly techniques through the thickness involved in a finite element implementation, such as Reddy et al. ([7], [8]). However, the use of Lagrange multipliers is preferred herein in order to directly assess the interlaminar transverse stresses.

The principle associated to the functional (8) is commonly called the first modified principle of potential energy with relaxed continuity requirements because the displacement functions in each element (or layer, in this case) may be chosen independently without any concern about the conformity requirement [9]. This approach can be interpreted as a hybrid formulation in the thickness.

In the analysis, each lamina is treated as a homogenous, orthotropic, elastic body whose behaviour is described by the Hooke’s Law,

\begin{align}
\sigma_{AB}^{(i)} &= C_{ABC}^{(i)} \varepsilon_{C}^{(i)} + C_{AB33}^{(i)} \varepsilon_{33}^{(i)} \\
\sigma_{A3}^{(i)} &= C_{A333}^{(i)} \varepsilon_{A3}^{(i)} \\
\sigma_{33}^{(i)} &= C_{3333}^{(i)} \varepsilon_{33}^{(i)}
\end{align}

in which $C_{\alpha\beta\gamma\eta}$ are the Cartesian components of the fourth-order Constitutive tensor.

On substituting of (1), (7) and (10) into (8), making some trivial manipulations, and integrating by parts at layer level (details are omitted for brevity), the stationary conditions is,

\begin{align}
\delta \Pi_{mP} = \sum_{i=1}^{n} \left\{ \int_{\Omega_{i}} \left( \delta u_{A}^{(i)} E_{i}^{(i)} C_{ABCD}^{(i)} \Delta_{D}^{(i)} + \delta u_{A}^{(i)} E_{ikl}^{(i)} C_{AB33}^{(i)} \Delta_{kl}^{(i)} + \delta u_{A}^{(i)} E_{k}^{(i)} C_{\alpha3}^{(i)} \Delta_{\alpha}^{(i)} + \delta u_{A}^{(i)} E_{kkl}^{(i)} C_{\alpha33}^{(i)} \Delta_{\alpha k}^{(i)} + \delta u_{A}^{(i)} E_{kl}^{(i)} C_{\alpha333}^{(i)} \Delta_{\alpha l}^{(i)} + \delta u_{A}^{(i)} E_{kkl}^{(i)} C_{\alpha3333}^{(i)} \Delta_{\alpha l}^{(i)} \right) d\Omega \right\} + \\
+ \sum_{j=1}^{i} \left\{ \int_{\Sigma_{j+1}} \left( s_{\alpha}^{(i)} \delta u_{\alpha}^{(i)} - s_{\alpha}^{(i)} \delta u_{ab}^{(i)} \right) dS + \delta W^{\text{ext}} \right\} = 0
\end{align}

where $\delta$ means the variational operator, and the integration over domain is represented as,
Integration across the thickness is carried out explicitly due to the displacement field in the transverse coordinate, reducing the three-dimensional problem to a two-dimensional one, as usual in plate theory. To the purpose, the following definitions have been introduced,

\[ (E_{k,l}^{(i)}, E_{k,l}^{(i)}, E_{k,l}^{(i)}, E_{k,l}^{(i)}) = \int_{h_i} (F_k F_l F_{k,l}, F_{k,l}, F_k, F_{k,l}) \, dz \]  

(13)

Since the variations are arbitrary, we obtain the \((9n+3)\) Euler-Lagrange equations that governing the problem,

\[ \delta u_{A_1}^{(i)} : \quad E_{t,i}^{(i)} C_{A,B,C,D} u_{l,D}^{(i)} + E_{t,j}^{(i)} C_{A,B} u_{l,D}^{(i)} + E_{t,j}^{(i)} C_{A,C} u_{j}^{(i)} + E_{t,j}^{(i)} C_{A,D} u_{l}^{(i)} + s_A^{(i)} + p_A^{(i)} \]  

(14)

\[ \delta u_{A_2}^{(i)} : \quad E_{b,i}^{(i)} C_{A,B,C,D} u_{l,D}^{(i)} + E_{b,j}^{(i)} C_{A,B} u_{l,D}^{(i)} + E_{b,j}^{(i)} C_{A,C} u_{j}^{(i)} + E_{b,j}^{(i)} C_{A,D} u_{l}^{(i)} + s_A^{(i+1)} + p_A^{(i)} \]  

(15)

\[ \delta u_{A_3}^{(i)} : \quad E_{b,i}^{(i)} C_{A,B,C,D} u_{l,D}^{(i)} + E_{b,j}^{(i)} C_{A,B} u_{l,D}^{(i)} + E_{b,j}^{(i)} C_{A,C} u_{j}^{(i)} + E_{b,j}^{(i)} C_{A,D} u_{l}^{(i)} + s_A^{(i+1)} = p_A^{(i)} \]  

(16)

\[ \delta S_{A_2}^{(j)} : \quad u_{l}^{(j)} - u_{l}^{(j-1)} = 0 \]  

(18)

Explicit form of the governing equations for each layer can be written by expanding the introduced subscripts and superscripts over their ranges. The boundary conditions are omitted for brevity. Set of equations (14)-(18) describes \((6n)\) second-order Partial Differential Equations (PDE) coupling to \((3n-3)\) algebraic equations in terms of \((9n-3)\) variables.

3 FINITE ELEMENT DISCRETIZATION

The weak form of the equations that governing the problem are obtained multiplying the Euler-Lagrange equations by their respective variations, integrating over the element domain, and integrating by parts to weaken the differentiability of \(\delta u_{A_k}^{(i)}(\alpha=1,2,3; k=t,b)\),

\[ \int_{D_i} \left( \delta u_{A_k}^{(i)} E_{t,i}^{(i)} C_{A,B,C,D} u_{l,D}^{(i)} + \delta u_{A_k}^{(i)} E_{t,j}^{(i)} C_{A,B} u_{l,D}^{(i)} + \delta u_{A_k}^{(i)} E_{t,j}^{(i)} C_{A,C} u_{j}^{(i)} + \delta u_{A_k}^{(i)} C_{A,D} u_{l}^{(i)} + s_A^{(i)} + p_A^{(i)} \right) \, d\Omega = 0 \]  

(19)

\[ \int_{D_i} \left( \delta u_{A_k}^{(i)} E_{b,i}^{(i)} C_{A,B,C,D} u_{l,D}^{(i)} + \delta u_{A_k}^{(i)} E_{b,j}^{(i)} C_{A,B} u_{l,D}^{(i)} + \delta u_{A_k}^{(i)} E_{b,j}^{(i)} C_{A,C} u_{j}^{(i)} + \delta u_{A_k}^{(i)} C_{A,D} u_{l}^{(i)} + s_A^{(i)} + p_A^{(i)} \right) \, d\Omega = 0 \]  

(20)
where the notation $\delta(*)|(.)$ has been introduced to denote the variation of (*) with respect to (\).

The kinematic variables require $C^0$ continuity in-plane due to just their first derivative are present in the expressions (19)-(23), while the statics variables can be discontinuous ($C^1$ Continuity). However, by simplicity the same interpolation is used for every unknown.

Using isoparametric finite element [10],

$$u_{\alpha k}^{(i)} = \sum_{m=1}^{NN} N_m^{(i)} \tilde{u}_{\alpha k}^{m} \quad \text{for } \alpha = t, b; \ i = 1, ..., n \quad (24a)$$

$$s_{\alpha}^{(i)} = \sum_{m=1}^{NN} N_m^{(i)} \tilde{s}_{\alpha}^{m} \quad \text{for } \alpha = 2, ..., n \quad (24b)$$

$$x_{\alpha} = \sum_{m=1}^{NN} N_m^{(i)} \tilde{x}_{\alpha}^{m} \quad (24c)$$

in which $N_m(\xi, \eta)$ are the shape functions associated to $m^{th}$-node and $NN$ is the number of nodes of the element (in this paper, a four-node quadrilateral element is used, i.e., $NN=4$). The circumflex tilde denotes nodal variable.

As is well known, the four-node quadrilateral element with transverse shear strain locks when it is thin (assuming full numerical integration) [10]. In order to avoid the shear-locking, the mixed interpolation of tensorial components (MITC, see [11]) technique is applied to assess the transverse shear strain at layer level, resulting,

$$\gamma_{13} = \frac{C_{y} + \frac{E_{y} t}{\text{det}(J)}}{8} \left( 1 + \eta \right) \left\{ F_{t} \left[ \tilde{g}_{1}^{(i)} - \tilde{g}_{2}^{(i)} \right] + F_{b} \left[ \frac{\tilde{g}_{1}^{(i)} - \tilde{g}_{2}^{(i)}}{2} \right] - \frac{\tilde{x}_{1}^{(i)} - \tilde{x}_{2}^{(i)}}{4h_i} \left( \tilde{u}_{1}^{(i)} + \tilde{u}_{2}^{(i)} \right) \right\} +$$

$$- \frac{\tilde{x}_{1}^{(i)} - \tilde{x}_{2}^{(i)}}{4h_i} \left( \tilde{u}_{1}^{(i)} + \tilde{u}_{2}^{(i)} \right) + \frac{\tilde{x}_{1}^{(i)} - \tilde{x}_{2}^{(i)}}{4h_i} \left( \tilde{u}_{1}^{(i)} + \tilde{u}_{2}^{(i)} \right) + \frac{\tilde{x}_{1}^{(i)} - \tilde{x}_{2}^{(i)}}{4h_i} \left( \tilde{u}_{1}^{(i)} + \tilde{u}_{2}^{(i)} \right) +$$

$$+ (1 - \eta) \left\{ F_{t} \left[ \tilde{g}_{1}^{(i)} - \tilde{g}_{2}^{(i)} \right] + F_{b} \left[ \frac{\tilde{g}_{1}^{(i)} - \tilde{g}_{2}^{(i)}}{2} \right] - \frac{\tilde{x}_{1}^{(i)} - \tilde{x}_{2}^{(i)}}{4h_i} \left( \tilde{u}_{1}^{(i)} + \tilde{u}_{2}^{(i)} \right) \right\} +$$

$$- \frac{\tilde{x}_{1}^{(i)} - \tilde{x}_{2}^{(i)}}{4h_i} \left( \tilde{u}_{1}^{(i)} + \tilde{u}_{2}^{(i)} \right) + \frac{\tilde{x}_{1}^{(i)} - \tilde{x}_{2}^{(i)}}{4h_i} \left( \tilde{u}_{1}^{(i)} + \tilde{u}_{2}^{(i)} \right) + \frac{\tilde{x}_{1}^{(i)} - \tilde{x}_{2}^{(i)}}{4h_i} \left( \tilde{u}_{1}^{(i)} + \tilde{u}_{2}^{(i)} \right) +$$

$$- \frac{\tilde{x}_{1}^{(i)} - \tilde{x}_{2}^{(i)}}{4h_i} \left( \tilde{u}_{1}^{(i)} + \tilde{u}_{2}^{(i)} \right) + \frac{\tilde{x}_{1}^{(i)} - \tilde{x}_{2}^{(i)}}{4h_i} \left( \tilde{u}_{1}^{(i)} + \tilde{u}_{2}^{(i)} \right)$$

$$(25a)$$
The determinant of the Jacobian matrix is,

\[
\det(f) = \frac{A_x + \xi B_x}{\det(f)} \left( 1 + \eta \right) \left[ F_t \theta_1^0 + \frac{\theta_1^0 - \theta_2^0}{2} + F_b \frac{\theta_1^0 - \theta_2^0 - \theta_3^0}{4 h} \left( \hat{u}_1^0 + \hat{u}_2^0 \right) + \frac{\theta_1^0 - \theta_2^0 - \theta_3^0}{4 h} \left( \hat{u}_2^0 + \hat{u}_3^0 \right) \right] + \frac{\theta_1^0 - \theta_2^0 - \theta_3^0}{4 h} \left( \hat{u}_1^0 + \hat{u}_2^0 \right)
\]

and

\[
A_x = \xi^1 - \xi^2 - \xi^3 + \xi^4 \quad B_x = \xi^1 - \xi^2 + \xi^3 - \xi^4 \quad C_x = \xi^1 + \xi^2 - \xi^3 - \xi^4 \quad A_y = \xi^1 - \xi^2 \quad B_y = \xi^1 + \xi^2 + \xi^3 - \xi^4 \quad C_y = \xi^1 + \xi^2 - \xi^3 - \xi^4
\]

Introducing (24) and (25) into (19)-(23), after some manipulations, the finite element model is obtained. For a three-layered plate,

\[
\begin{bmatrix}
K_{tt}^{(i)} & K_{tb}^{(i)} & 0 \\
K_{tb}^{(i)} & K_{bb}^{(i)} & -Q_{lb} \\
0 & -Q_{lb} & 0
\end{bmatrix}
\begin{bmatrix}
\hat{u}_t^{(i)} \\
\hat{u}_b^{(i)} \\
\hat{s}^{(2)}
\end{bmatrix}
= \begin{bmatrix}
F_t^{(i)} \\
F_b^{(i)} \\
0
\end{bmatrix}
\]

where \(K_{tt}^{(i)}\), \(K_{tb}^{(i)}\) and \(K_{bb}^{(i)}\) are the stiffness matrices related to top-top, top-bottom and bottom-bottom degrees of freedom, respectively, and \(Q_{lb}\) is the matrix resulting from the Lagrangian constraint problem. The arrays \((F_t^{(i)}, F_b^{(i)})\) represent the equivalent nodal loads applied at the \(i^{th}\)-layer. Explicit form of mentioned arrays could be obtained as is common in finite element technique; herein, these are omitted by brevity.

Static variables (interlaminar transverse stresses) can be computed directly eliminating the displacement variables. After some algebra, we can write,
where,

\[
\begin{align*}
\mathbf{R}_t^{(i)} &= \mathbf{K}^{(i)}_{tt} - \mathbf{K}^{(i)}_{tb} \cdot \mathbf{K}^{(i)}_{bb}^{-1} \cdot \mathbf{K}^{(i)}_{tb} \quad (30) \\
\mathbf{R}_b^{(i)} &= \mathbf{K}^{(i)}_{tt} - \mathbf{K}^{(i)}_{tb} \cdot \mathbf{K}^{(i)}_{tb}^{-1} \cdot \mathbf{K}^{(i)}_{tb} \quad (31) \\
\mathbf{R}_b^{(i)} &= \mathbf{R}_t^{(i)} \cdot \mathbf{R}_b^{(i)} \quad (32) \\
\mathbf{F}^{(i)} &= \mathbf{Q}_{tt}^{(i-1)} \cdot \mathbf{K}^{-1}_{tt} \cdot \mathbf{F}_{t}^{(i-1)} - \mathbf{Q}_{tb}^{(i-1)} \cdot \mathbf{K}^{-1}_{tb} \cdot \mathbf{Q}_{tb}^{(i-1)} \cdot \mathbf{F}_{b}^{(i-1)} + \\
&\quad - \mathbf{Q}_{tb}^{(i)} \cdot \mathbf{K}^{-1}_{tb} \cdot \mathbf{F}_{t}^{(i)} + \mathbf{Q}_{tb}^{(i)} \cdot \mathbf{R}_{tb}^{(i)} \cdot \mathbf{F}_{b}^{(i)} \quad i = 2, ..., n \quad (33)
\end{align*}
\]

The obtained matrix (29) is both symmetric and positive-definite. Necessary but not sufficient stability condition to solve it is [12],

\[
n_u \geq n_s \quad (34)
\]

in which \(n_u\) and \(n_s\) are the total number of degrees of freedom associated with the kinematic and static variables, respectively.

Note that, if desired, the displacements could be subsequently determined from the static variables.

4 NUMERICAL RESULTS

The first example is a free-edge bending problem solved by Tahani and Nosier [13]. Consider a \([0°/90°]\), cross-ply laminate subjected to uniform transverse load \(q_o\) on its top surface. The load is applied downward. The geometric of the plate, simply supported on two opposite edge and free in the other remaining two, meets the following relationships: \(a=10h\) and \(b=4h\), where \(a,b\) are the dimensions coincident with the \(x,y\) directions, respectively, and \(h\) is the total thickness (Figure 1). Each lamina is assumed to be of the same thickness \(h_i\) and has the following material properties,

\[
\begin{align*}
E_L &= 20 \times 10^6 \text{ psi}; \quad E_T = 2.1 \times 10^6 \text{ psi}; \\
G_{LT} &= G_{TT} = 0.85 \times 10^6 \text{ psi}; \quad v_{LT} = v_{TT} = 0.21
\end{align*}
\]

(35)

where the subscripts L,T signify the direction parallel to the fibers and the transverse direction perpendicular to the fibers, respectively. Due to the symmetry, just one quarter of the plate is modelled. The in-plane mesh has 5×11 four-node quadrilateral elements. All elements have the same length (i.e., \(a/2/5\)). However, a fine finite element mesh near the free edges was used in the modelling (see Figure 3). Thus, the width of the elements decreases as the free edge at \((x,b/2,z)\) is approached. The widths of the eleven rows of elements, as one moves away from the refined free edge, are \(h_i/32, h_i/32, h_i/16, h_i/8, h_i/4, h_i/2, h_i, 1.5h_i, 1.5h_i, 2.5h_i\). By the same criterion, the thickness of the numerical layers decreases as the interface between two physical layers is approached. From the top to the midplane, the layer thickness are \(h_i/3, h_i/3, h_i/6, h_i/6, h_i/6, h_i/6, h_i, h_i/2, h_i, h_i, h_i/2, h_i\). The boundary conditions applied are,

\[
\begin{align*}
(0,y,z): \quad u_1 = u_3 = 0; \quad (a/2,y,z): \quad u_1 = 0; \quad (x,0,z): \quad u_2 = 0
\end{align*}
\]

(36)
Figure 3: In-Plane mesh (left) and discretization across the thickness, on the \( yz \)-plane (right), used to model the free edge bending problem.

All the numerical results are normalized as follows,

\[
\bar{\sigma}_{A3} = \frac{h}{q_0} \sigma_{A3} \quad \text{(37a)}
\]

\[
\bar{\sigma}_{33} = \frac{1}{q_0} \sigma_{33} \quad \text{(37b)}
\]

Figure 4 illustrates the variation of transverse normal stress through the thickness at \( x=5h \). The distribution obtained is in good accordance with those by Tahani and Nosier [13]. The plot shown suggests that a stress singularity may exist at the intersections of the \( 0^\circ/90^\circ \) interfaces and the free-edges, where the maximum bending stress \( \sigma_{33} = q_0 \) is approximately tripled. However, such a conjecture cannot be proved by the use of approximate solutions such as numerical methods. This interlaminar stress concentration near the edge of the laminate can initiate delamination damage resulting in a significant loss of stiffness.

Figure 4: Normalized transverse normal stress \( \sigma_{33} \) through the thickness at \( x=5h \) and the free-edge. Laminate \([0^\circ/90^\circ]_s\). Comparison with solution of Tahani & Nosier [14].
Typical distributions of the interlaminar shear stress $\sigma_{23}$ across the middle plane ($z=0$) and the upper $0^\circ/90^\circ$ interface ($z=-h/4$) laminate at $x=2h$ are shown in Figure 5. It is noted that the magnitude of $\sigma_{23}$ is increased in the boundary layer region and exhibits singular behaviour near the free edge at $z=h/4$. It is seen that the numerical value of $\sigma_{23}$ may not become zero at the free edge, contrary to what is expected. Whitcomb et. al.\cite{14} suggest that this behaviour is due to the by non-symmetry of the stress tensor at a singular point.

![Figure 5: Distributions of transverse shear stress $\sigma_{23}$ throughout the width at $z=0$ and $z=-h/4$. Laminate $[0^\circ/90^\circ]_2$.](image)

Based on the previous problem, the effect of the lamination scheme is studied. Figure 6 shows the variation of the transverse normal stress at $x=0.50h$ as the free edge is approached for $[0^\circ/90^\circ]_2$ and $[90^\circ/0^\circ]_2$ laminates.

![Figure 6: Normalized transverse normal stress $\sigma_{33}$ through the thickness at $x=0.50h$. Left: Laminate $[0^\circ/90^\circ]_2$. Right: Laminate $[90^\circ/0^\circ]_2$.](image)

Because of the different position of the rigid $0^\circ$ layer, a different bending behaviour is
expected. It’s noted that the maximum magnitude of $\sigma_{33}$ in the $[0^\circ/90^\circ]$ lay-up is greater than that of the $[90^\circ/0^\circ]$, as expected for the highest rigidity scheme. Also, while in the first case the $\sigma_{33}$ peaks are compressive, in the second one are tensile. Further away from the free edge, peaks disappear and the transverse normal stress is completely compressive in both cases. It is seen that the transverse normal stress profile near the free edge is strongly influenced by the stacking sequence.

5 CONCLUSIONS

A layer-wise mixed finite element model was presented to evaluate the interlaminar transverse stresses directly without any post-processing technique by means Lagrange multipliers. Several examples have been carried out. It is shown that the obtained results are very close to those available in the literature. Further, the lay-up sensibility was study, concluding that the stacking sequence has a high influence on the free-edge effect in bending.

REFERENCES

FREE VIBRATIONS OF AN AXIALLY FUNCTIONALLY GRADED BEAM WITH A CONCENTRATED MASS

GONZALO J. GILARDI¹, DIANA V. BAMBILL¹ AND CARLOS A. ROSSIT¹

¹ Departamento de Ingeniería – Instituto de Mecánica Aplicada (IMA) Universidad Nacional del Sur
Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)
Avenida Alem 1253 – (8000) Bahía Blanca - Argentina
gilardigonzalo@gmail.com
dbambill@criba.edu.ar
carossit@criba.edu.ar

Key words: Beam vibrations, Functionally graded material, Axially functionally graded, attached mass.

Abstract. The free transverse vibrations of axially functionally graded (AFG) cantilever beams with a concentrated mass at the free end are studied in this paper. The material properties of the AFG beam, consisting of metal and ceramic, vary continuously in the axial direction according to a power-law form. Approximated solutions for the title problem are obtained by means of the Ritz Method. The influence of the material variation on the natural frequencies of vibration of the functionally graded beam is investigated and compared with the influence of the variation of the cross section. The phenomenon of dynamic stiffening of beams can be observed in various situations. Through results available in the literature that can be represented by the model under study, the accuracy of the procedure is verified.

1 INTRODUCTION

The situation of structural elements supporting motors or engines attached to them is usual in technological applications. The operation of the attached machine may introduce severe dynamic stresses on the beam. It is important, then, to predict the natural frequencies of the coupled beam-mass system, in order to obtain a proper design of the structural elements.

Beams are probably the most widespread component of resistant structures. Consequently, enormous research effort has been developed about them. In recent years, it has become increasingly important to use advanced materials whose properties vary gradually in some of its dimensions (FGM).

In present paper, beams of materials whose properties vary functionally along the axis (AFG) and the cross section varies gradually (tapered) are studied.

Such materials were first used by Japanese researchers to the mid-eighties, Niino et al.[1].
They constitute a class of advanced materials whose properties vary gradually with some of its dimensions. Different from composite materials, in FGM materials abrupt interfaces are removed and replaced with a graded interface that produces a smooth transition between the material and next. One of their distinguishing features is the ability to adapt to a specific application.

These materials can be manufactured by varying the percentage content of two or more materials so that the new material has a property such as Young’s modulus, material density, etc with a desired gradient in established spatial directions.

Thus, thermal stress, residual stress and stress concentration factors found in composite laminates are reduced

Under the favorable characteristics mentioned above, there have been many studies on the static and dynamic behavior of resistant structures of FGM.

Initially, investigative work established great progress in the field of elasticity theory and the study of plates and shells built with FGM. Paradoxically progress in its application to beams (Functionally Graded Beams -FGB-) was much later.

Sankar [2] established a model of FG Euler-Bernoulli beam to treat a static problem in a simply supported beam. The corresponding case of free transverse vibrations was studied by Aydogdu and Taskin [3]. Additionally Chabraborty et al. [4] developed a finite element to study the thermoelastic behavior of FGB. In the works of Lu and Chen[5] and Zhong and Yu [6] complex analytical procedures are carried out to solve accurately FGB within the field of the Theory of Elasticity. All these studies were conducted based on the Euler-Bernoulli description of the beam deformation. As is known the classical Bernoulli-Euler beam theory adequately predicts the frequencies of vibration of lower modes of slender beams

A survey of the literature reveals that most of the early work on functionally graded beams has considered the gradation of the material properties in the thickness direction. Far fewer researchers have considered the variation of material properties in the axial direction (AFG). Probably because the problem becomes more complicated as variable coefficients appear in the governing equations.

Consequently, because of the mathematical difficulties in treatment, few analytical solutions have been obtained and for arbitrary specific gradients: For example Elishakoff and his colleagues [7-10], by means of the semi-inverse method, solved various particular cases of vibrating AFG beams.

Huang and Li  [11] solved the problem by transforming the governing equation with variable coefficients in a Fredholm integral equation


Akgöz and Civalek [17], studied the free longitudinal vibrations of AFG bars on the basis of strain gradient elasticity theory by using the Rayleigh-Ritz method.

A recent literature survey on free vibration of functionally graded beams revealed that not many papers consider the influence of an attached mass. Ramesh and Rao [18], studied the natural frequencies of a rotating FGM beam with a concentrated mass. The material properties
of the FGM beam symmetrically vary continuously in thickness direction.
In particular to the authors’ knowledge [19], there are no natural frequency data in the
literature for axially functionally graded beams, AFG, carrying concentrated masses.
In this paper, natural vibrations of an AFG cantilever beam with a tip mass at the free end
are studied. The beam cross section may also have variable height \( h(x) \) and variable width
\( b(x) \). The Ritz method [20] is advantageously employed to solve the problem.

2 THEORY

2.1 Axially functionally graded material properties

![Power law relation of AFG material properties.](image)

The inhomogeneous material, with gradient compositional variation of the constituents,
varies in the longitudinal direction of the beam. Properties of AFG materials, like mass
density \( \rho \), Young’s modulus \( E \), shear modulus \( G \), continuously vary in the axial direction.

For example, a generic material property \( P(\bar{x}) \) is assumed to vary along the beam axis \( \bar{x} \)
with a power law relation, Fig. 1.:

\[
P(\bar{x}) = P_a + (P_b - P_a) \left( \frac{\bar{x}}{L} \right)^n
\]

where \( P_a \) and \( P_b \) are properties of material “a” and material “b”, respectively. They are the
constituents of the inhomogeneous material of the beam; \( n \) is the material non-homogeneity
parameter and \( P(\bar{x}) \) is a typical material property such as \( \rho \), \( E \) or \( G \). Note that for \( x = 0 \) the
entire section is of material “a”, and for \( x = L \) the whole section is of material “b”. The
percentage content of material “a” along the beam increases as \( n \) increases. When \( n=1 \) the composition changes linearly through the length \( L \), while \( n=1/2 \) or \( n=2 \) corresponds to a quadratic distribution, and so on. In general, any value \( n \) outside the range \((1/3, 3)\) is not desired [21] because such a functionally graded material would contain too much of one of the constituents. (When \( n=1/3 \) or 3, one constituent has the 75% of the total AFG material).

2.2 Analytical procedure

As shown in Fig. 2, the considered model is a cantilever AFG beam with variable cross section and carrying a concentrated masses \( m \) located at \( x=L \).

In order to find the natural frequencies of the system one assumes that the beam deflection \( v(\bar{x},t) \) may be expressed in the form

\[
v(\bar{x},t) = V(\bar{x}) \cos(\omega t)
\]

where \( \omega \) is the natural circular frequency.

To apply the Rayleigh Ritz method, it is necessary to approximate the spatial component of the solution:

\[
V(x) \cong V_a(x) = \sum_{j=1}^{p} C_j \phi_j
\]

\( \phi_j \) are coordinate functions that satisfy the essential boundary conditions, \( C_j \) are arbitrary constants.

The following non-dimensional coordinate is introduced:
\[ x = \frac{x}{l} \]  

(4)

For the problem in analysis, a cantilever beam, the following coordinate functions are chosen:

\[ \{ \varphi \}_j = \{ x^{(3j)} \}_j \]

(5)

which satisfy the essential boundary conditions:
\[ \varphi(0) = \varphi'(0) = 0 \]

The energy functional \( J \) for beam vibration problem can be written in terms of the maximum strain energy and the maximum kinetic energy of the beam-mass system as

\[ J = U_{\text{max}} - T_{\text{max}} \]

(6)

where, for an Euler-Bernoulli beam and when the rotatory inertia of the concentrated mass is taken into account:

\[ U_{\text{max}} = \frac{1}{2} \int_0^l E(x) I(x) \left[ \frac{d^2 V_a}{dx^2} \right]^2 dx \]

(7)

and

\[ T_{\text{max}} = \frac{1}{2} \omega^2 \left\{ \int_0^l \rho(x) A(x) L [V_a(x)]^2 dx + m [V_a'(1)]^2 + \frac{m r_g^2}{L^2} \left[ \frac{dV_a(x)}{dx} \right]^2 (1) \right\} \]

(8)

where \( r_g \) is the radius of gyration of the mass with respect to the neutral axis of the beam.

As the material and geometric characteristics of the beam may be general, one can define:

\[ E(x) = E_0 f_E(x) \]
\[ I(x) = I_0 f_I(x) \]
\[ \rho(x) = \rho_0 f_\rho(x) \]
\[ A(x) = A_0 f_A(x) \]
\[ b(x) = b_0 f_b(x) \]
\[ h(x) = h_0 f_h(x) \]

Obviously, \( f_A = f_b \times f_h, \quad f_I = f_b \times f_h^3 \)

Replacing (7) and (8) into (6) and taking into account (9), the functional can be expressed:
\[ J(V_a) = \frac{1}{2} \frac{E_0 I_0}{L^3} \times \]
\[ \left\{ \int_0^1 f_x f_t \left[ \frac{d^2 V_a}{dx^2} \right]^2 dx - \Omega^2 \left[ \int_0^1 f_x f_t L [V_a(x)]^2 dx + M \left[ \frac{dV_a(1)}{dx} \right]^2 + c_2 \left[ \frac{dV_a(x)}{dx} \right]^2 \right] \right\} \quad (10) \]

with: \( \Omega = \omega l^2 \frac{\rho_0 A_h}{E_0 I_0} \), \( M = \frac{m}{\rho_0 A_h l} \), \( c = \frac{r_x}{l} \)

Following Ritz’ procedure, the functional is minimizing with respect to every arbitrary constant:
\[ \frac{\partial J[V_a(x)]}{\partial C_j} = 0 \quad j = 1, 2, ..., P \quad (11) \]

Then a linear system of equations is formed:
\[ A \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_P \end{bmatrix} = 0 \quad (12) \]

which results in the following eigenvalue equation:
\[ A = \Omega^2 M \quad (13) \]

where
\[ k_{ij} = \int_0^1 f_x(x) f_t(x) \phi_i \phi_j dx \quad (14) \]
\[ m_{ij} = \int_0^1 f_x(x) f_t(x) \phi_i \phi_j dx + M \left[ \phi_i(1) \phi_j(1) + c^2 \phi_i(1) \phi_j(1) \right] \quad (15) \]

\( k_{ij} \) and \( m_{ij} \) are the elements of matrices \( K \) and \( M \), respectively

Then, the eigenvalue problem can be expressed as:
\[ \left| K M^{-1} - \Omega^2 I \right| = \left| B - \lambda I \right| = 0 \quad (16) \]

where \( \lambda = \sqrt{\Omega} \) are the eigenvalues of matrix \( B \)
3 NUMERICAL RESULTS

Since there were not found, in the technical literature, values of natural frequencies of vibration of AFG beams with attached masses in order to verify the accuracy of the proposed model, comparisons are made with particular cases available in the literature.

First, Table 1 compares values for a tapered Euler-Bernoulli beam made of axially functionally graded material studied by Shahba and Rajasekaran [13]. They obtained values for the first two natural frequency coefficients for a case that can be represented in the present model by adopting in Equations (9):

\[ f_b = 1 - c_b x \]
\[ f_h = 1 - c_h x \]
\[ f_E = 1 + x \]
\[ f_{\rho} = 1 + x + x^2 \]

The calculations were done with \( p = 20 \) in equation (3) in all cases.

<table>
<thead>
<tr>
<th>( C_b )</th>
<th>( \Omega_1 )</th>
<th>( \Omega_2 )</th>
<th>( \Omega_1 )</th>
<th>( \Omega_2 )</th>
<th>( \Omega_1 )</th>
<th>( \Omega_2 )</th>
<th>( \Omega_1 )</th>
<th>( \Omega_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Present</td>
<td>2.42556</td>
<td>18.6041</td>
<td>2.60542</td>
<td>19.0041</td>
<td>2.85075</td>
<td>19.5303</td>
<td>3.21368</td>
</tr>
<tr>
<td>0.2</td>
<td>Present</td>
<td>2.50506</td>
<td>17.3802</td>
<td>2.68633</td>
<td>17.7501</td>
<td>2.93357</td>
<td>18.2379</td>
<td>3.29935</td>
</tr>
<tr>
<td>0.4</td>
<td>Present</td>
<td>2.61547</td>
<td>16.0705</td>
<td>2.79874</td>
<td>16.4092</td>
<td>3.04857</td>
<td>16.8571</td>
<td>3.41810</td>
</tr>
<tr>
<td>0.6</td>
<td>Present</td>
<td>2.78355</td>
<td>14.6508</td>
<td>2.96694</td>
<td>14.9567</td>
<td>3.22368</td>
<td>15.3627</td>
<td>3.59847</td>
</tr>
</tbody>
</table>

As it can be seen, the agreement is excellent.

In the second place, Table 2 shows the results for an homogeneous cantilever beam with
uniform cross section with a tip mass attached at the free end of the beam. In this case comparison is made with the paper by Rossit and Laura [22], where the eigenvalues $\lambda_i$ of the system are obtained.

Then, to obtain the frequency coefficients $\Omega_i$, one must square the eigenvalues. Again, the concordance is excellent

Table 2: Eigenvalues of a vibrating homogeneous cantilever beam with a mass $m$ attached at the free end

<table>
<thead>
<tr>
<th>$M$</th>
<th>Present</th>
<th>[22]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_1$</td>
<td>1.61640</td>
<td>1.6164</td>
</tr>
<tr>
<td>$\lambda_2$</td>
<td>4.26706</td>
<td>4.2671</td>
</tr>
<tr>
<td>$\lambda_3$</td>
<td>7.31837</td>
<td>7.3184</td>
</tr>
<tr>
<td>$\lambda_4$</td>
<td>10.4016</td>
<td>10.4020</td>
</tr>
<tr>
<td>$\lambda_5$</td>
<td>13.5067</td>
<td>13.5070</td>
</tr>
</tbody>
</table>

Due to the quantity and variability of the parameters involved in the description of the dynamical behavior of these kinds of structures, just a few representative cases will be considered to demonstrate the convenience of the procedure.

In the calculations, the constituents of the inhomogeneous material are assumed to be steel and aluminum oxide $\text{Al}_2\text{O}_3$ (alumina). Their Young modulus and density are:

- $E_\text{St} = 210 \text{GPa}$; $\rho_\text{St} = 7800 \text{kg/m}^3$;
- $E_\text{Al} = 390 \text{GPa}$; $\rho_\text{Al} = 3960 \text{kg/m}^3$;
- $v_\text{St} = v_\text{Al} = 0.30$

The relationships between material properties are:

- $E_\text{Al}/E_\text{St} = 1.857$ for Young's modulus
- $\rho_\text{Al}/\rho_\text{St} = 0.508$ for the density. Note that the alumina, more rigid, is lighter than steel.

In order to evaluate the influence of the material composition, an AFG beam of uniform cross section with a mass at the free end is analyzed.

A linear distribution of the materials is considered - $n=1$ in Eq.(1) - and two possibilities are explored:

a) Material $a$ in Eq.1 is Steel and material $b$ is Alumina.

Then, in $x=0$, the section is of Steel and in $x=L$ the section is entirely of Alumina

b) Material $a$ in Eq.1 is Alumina and material $b$ is Steel.

Then, in $x=0$, the section is of Alumina and in $x=L$ the section is entirely of Steel
The use of such materials in the case of beams that support electromechanical equipment, represented by the mass, is concerned in order to dynamically stiffen the beam. The goal is to raise its natural frequencies, mainly the fundamental frequency, away from operating frequency of the attached equipment. Thus resonance phenomenon is avoided.

For this, it is intended that alumina, stronger and lighter, predominates in areas of the beam with higher curvature (greater bending moment) and greater displacement. Thus the rigidity of the material increases the strain energy and its reduced weight decreases the inertial effect. Consequently, natural frequencies increase.

This is achieved very simply in beams with ends prevented to displace (simply supported, clamped) where the areas of greater curvature and greater displacement converge around the central area of the beam [19, 23].

But that is not the case of cantilever beams, where the maximum curvature areas are around the clamped end and maximum displacement around the free end.

In Table 3, it can be seen that for the bare cantilever beam ($M=0$) values are almost the same, regardless of the order in which the constituent materials of the FG material are arranged. The values quoted in the table correspond to $\Omega_i = \omega_i^2 \frac{I^2}{\rho \bar{A}}$ in all cases

The situation is clearly different when the beam carries a mass at its free end. The fundamental frequency increases significantly with composition b). This is because the increase in bending moment by the weight of the mass exceeds the increase of the kinetic energy due to its inertial effect. For $M=0.5$ the increment in $\Omega_1$ is 38% and for $M=1$ is 48%.

Table 3: Frequency coefficients $\Omega_i$ of an AFG cantilever beam with a mass $m$ attached at the free end

<table>
<thead>
<tr>
<th>$M$</th>
<th>$c$</th>
<th>Material variation</th>
<th>$\Omega_1$</th>
<th>$\Omega_2$</th>
<th>$\Omega_3$</th>
<th>$\Omega_4$</th>
<th>$\Omega_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>a) St-Al</td>
<td>4.84848</td>
<td>30.1831</td>
<td>84.4216</td>
<td>165.366</td>
<td>273.306</td>
</tr>
<tr>
<td></td>
<td></td>
<td>b) Al-St</td>
<td>4.78429</td>
<td>30.0155</td>
<td>84.1921</td>
<td>165.098</td>
<td>273.015</td>
</tr>
<tr>
<td>0.5</td>
<td>0.1</td>
<td>a) St-Al</td>
<td>2.33166</td>
<td>19.7077</td>
<td>49.5917</td>
<td>95.0656</td>
<td>171.892</td>
</tr>
<tr>
<td></td>
<td></td>
<td>b) Al-St</td>
<td>3.22283</td>
<td>21.5339</td>
<td>53.9049</td>
<td>100.586</td>
<td>177.081</td>
</tr>
<tr>
<td>1</td>
<td>0.2</td>
<td>a) St-Al</td>
<td>1.70649</td>
<td>11.4947</td>
<td>34.3958</td>
<td>86.4294</td>
<td>166.878</td>
</tr>
<tr>
<td></td>
<td></td>
<td>b) Al-St</td>
<td>2.52282</td>
<td>13.1123</td>
<td>36.6694</td>
<td>88.8074</td>
<td>169.332</td>
</tr>
</tbody>
</table>

Then, the effect of the variation in cross section is evaluated. The width of the section $b$ is taken constant and variable height considering $f_h = 1 + \left( \frac{h_f}{h_0} - 1 \right) x^n$ in Eq. (9). Three values for $n$ were adopted: 0.5, 1(linear) and 2.

The obtained values are shown in Table 4. As can be seen the value $n = 2$ for the tapered beam is the one that produces a greater stiffening of the beam by raising its natural frequency.

418
Table 4: Frequency coefficients $\Omega_i$ of different homogeneous tapered cantilever beam with a mass $m$ attached at the free end

<table>
<thead>
<tr>
<th>$n$</th>
<th>$h_L/h_0$</th>
<th>$M$</th>
<th>$c$</th>
<th>$\Omega_1$</th>
<th>$\Omega_2$</th>
<th>$\Omega_3$</th>
<th>$\Omega_4$</th>
<th>$\Omega_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/2</td>
<td>0.2</td>
<td>0.5</td>
<td>0.1</td>
<td>0.89761</td>
<td>3.91033</td>
<td>10.8674</td>
<td>27.1203</td>
<td>51.7044</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>0.58944</td>
<td>1.60795</td>
<td>10.2991</td>
<td>26.8309</td>
<td>51.4948</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.5</td>
<td>0.1</td>
<td>1.25315</td>
<td>7.45078</td>
<td>16.7483</td>
<td>38.2350</td>
<td>72.1966</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>0.87456</td>
<td>3.32872</td>
<td>14.2801</td>
<td>37.0988</td>
<td>71.4352</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.5</td>
<td>0.1</td>
<td>1.53841</td>
<td>10.3033</td>
<td>23.4667</td>
<td>49.0560</td>
<td>91.1112</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>1.10769</td>
<td>5.15302</td>
<td>18.1300</td>
<td>46.4610</td>
<td>89.4484</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.5</td>
<td>0.1</td>
<td>1.78384</td>
<td>12.7530</td>
<td>30.5471</td>
<td>60.0981</td>
<td>109.465</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>1.31463</td>
<td>7.01143</td>
<td>22.0272</td>
<td>55.4379</td>
<td>106.553</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.2</td>
<td>0.5</td>
<td>0.1</td>
<td>1.33696</td>
<td>4.68944</td>
<td>12.7796</td>
<td>33.0209</td>
<td>63.9394</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>0.86190</td>
<td>1.95713</td>
<td>12.1359</td>
<td>32.6906</td>
<td>63.6969</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.5</td>
<td>0.1</td>
<td>1.65054</td>
<td>8.43140</td>
<td>18.3748</td>
<td>42.9372</td>
<td>81.9568</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>1.11949</td>
<td>3.73096</td>
<td>15.7597</td>
<td>41.7549</td>
<td>81.1530</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.5</td>
<td>0.1</td>
<td>1.77725</td>
<td>11.1235</td>
<td>24.7368</td>
<td>52.3270</td>
<td>97.8941</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>1.28118</td>
<td>5.1179</td>
<td>19.1703</td>
<td>49.6936</td>
<td>96.1961</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.5</td>
<td>0.1</td>
<td>1.90355</td>
<td>13.2154</td>
<td>31.2917</td>
<td>61.7901</td>
<td>112.965</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>1.40403</td>
<td>7.23201</td>
<td>22.5715</td>
<td>57.1066</td>
<td>110.034</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
<td>0.5</td>
<td>0.1</td>
<td>1.79995</td>
<td>5.65217</td>
<td>14.5755</td>
<td>39.1389</td>
<td>77.4703</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>1.18085</td>
<td>2.34140</td>
<td>13.7177</td>
<td>38.7087</td>
<td>77.1611</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.5</td>
<td>0.1</td>
<td>1.91157</td>
<td>9.57966</td>
<td>20.0263</td>
<td>47.6912</td>
<td>92.2991</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>1.35162</td>
<td>4.21289</td>
<td>17.0338</td>
<td>46.3822</td>
<td>91.4089</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.5</td>
<td>0.1</td>
<td>1.96190</td>
<td>12.0218</td>
<td>26.1118</td>
<td>55.5972</td>
<td>104.895</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>1.42547</td>
<td>5.94719</td>
<td>20.0687</td>
<td>52.8406</td>
<td>103.120</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.5</td>
<td>0.1</td>
<td>1.98827</td>
<td>13.6980</td>
<td>32.1192</td>
<td>63.4801</td>
<td>116.512</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>1.47165</td>
<td>7.49670</td>
<td>23.0459</td>
<td>58.7078</td>
<td>113.534</td>
<td></td>
</tr>
</tbody>
</table>

Finally, after the obtained results, the combination that yields the greatest dynamic stiffening for the scheme is analyzed: Height of the cross section varying parabolically with $n = 2$ and distribution b) of the material AFG.

As shown in Table 5 natural frequency values increase remarkably and in a widespread manner by incorporating the AFG material to the tapered beam.

The increase largely exceeds 60% in the first frequency for all situations, 40% for the second frequency and 30% in the rest of the frequencies calculated. It should also be pointed out that this increase in frequency is accompanied by a decrease in the weight of the beam.
Table 5: Frequency coefficients $\Omega$ of a tapered cantilever ($n=2$) AFG beam with a mass $m$ attached at the free end

<table>
<thead>
<tr>
<th>$h_L/h_0$</th>
<th>$M$</th>
<th>$c$</th>
<th>$\Omega_1$</th>
<th>$\Omega_2$</th>
<th>$\Omega_3$</th>
<th>$\Omega_4$</th>
<th>$\Omega_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.5</td>
<td>0.1</td>
<td>2.96570</td>
<td>8.11586</td>
<td>19.9883</td>
<td>52.2581</td>
<td>102.657</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.2</td>
<td>1.90478</td>
<td>3.53669</td>
<td>18.4773</td>
<td>51.4407</td>
<td>102.043</td>
</tr>
<tr>
<td>0.4</td>
<td>0.5</td>
<td>0.1</td>
<td>3.14880</td>
<td>13.6446</td>
<td>28.2017</td>
<td>64.8767</td>
<td>124.331</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.2</td>
<td>2.26732</td>
<td>6.20393</td>
<td>23.3396</td>
<td>62.5140</td>
<td>122.628</td>
</tr>
<tr>
<td>0.6</td>
<td>0.5</td>
<td>0.1</td>
<td>3.20885</td>
<td>17.1266</td>
<td>37.1762</td>
<td>76.7248</td>
<td>143.008</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.2</td>
<td>2.40216</td>
<td>8.75470</td>
<td>27.8263</td>
<td>71.9628</td>
<td>139.746</td>
</tr>
<tr>
<td>0.8</td>
<td>0.5</td>
<td>0.1</td>
<td>3.22608</td>
<td>19.5880</td>
<td>45.8881</td>
<td>88.5838</td>
<td>160.378</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.2</td>
<td>2.47675</td>
<td>11.05591</td>
<td>32.2431</td>
<td>80.6211</td>
<td>156.096</td>
</tr>
</tbody>
</table>

11 CONCLUSIONS

The advantageous use of FGM has been demonstrated in the design of a cantilever beam which supports a mass. The classical, variational method of Ritz has been successfully used to obtain an approximate, yet accurate, solution to a difficult elastodynamics problem from which data has not been found, by the authors, in the literature.

12 ACKNOWLEDGEMENTS

The present study has been sponsored by the Universidad Nacional del Sur, UNS, the Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET and the Comisión de Investigaciones Científicas, Buenos Aires Province (CIC). It was developed in the Departamento de Ingeniería and Instituto de Mecánica Aplicada, IMA.

REFERENCES


1 INTRODUCTION

Many engineering structures are composed by substructures assembled with bolted joints [1, 2]. Much research has been devoted to modeling the static behavior of these structures [3, 4] but research on the complex dynamics of bolted joints is scarce.

The dynamic behavior of structures with joints can be modeled with standard finite elements (such as beam elements), but these elements cannot reproduce the phenomenon of bolt looseness and friction between joint components which arises in real joints. Models of joints consisting of detailed 3D meshes have been developed [5, 6, 7], but they can be prohibitively expensive in terms of computational cost – especially with structures which incorporate many joints. Simplified joint models have been proposed [8, 9, 10]. Some of them include special nonlinear elements such as Jenkins elements [11] and models such as the Valanis and Bouc-Wen models [12, 13, 14]. These models have correctly reproduced energy dissipation under harmonic loading and have a very low computational cost [10]. However, its use is not extensive since commercial FE packages do not usually include them.

In this paper three simple computational models are presented towards an understanding the dynamic behavior of bolted joints with non-linear effects (see Figure 1). Simplified beam models were used to simulate the interaction between the bolt and the plates of a joint when the structure is excited with a harmonic load and the bolt is not fully tighten or even has intermittent contact (Cases 1 and 2, Figure 1), as well as the interaction between plates at areas between bolts (Case 3), incorporating the nonlinearities generated by the contact condition and friction. Due to these nonlinearities harmonic distortion arises in the analysis output. This phenomenon – well known in audio electronics – indicates that if the nonlinear system is excited with a mono frequency load, harmonics of that fundamental frequency appear at the output of the system. A heuristic formulation was developed for the prediction of the amplitude of the harmonics, given system physical parameters and excitation amplitude and frequency. This heuristic calculation could be applied in the future to develop simplified joint elements able to model bolt looseness and other nonlinear behavior.
2 GENERAL DESCRIPTION OF THE FINITE ELEMENT MODELS

Three 2-dimensional simplified models were developed with the commercial FE software ADINA V9.0 [15].

All models used linear, 2-noded Hermitian beams. A square cross section was used. An isotropic, linear elastic material formulation was used in all cases. Material properties are stated for each case.

Dynamic simulations were performed with the Newmark implicit direct integration method ($\delta = 0.5, \alpha = 0.5$), with a step size of $1E^{-4}$ s. When contact conditions were modeled, the Constraint Function method [16] was used. No damping was included.

3 DESCRIPTION OF THE SIMPLIFIED MODELS

3.1 Model I

This model is devised to analyze the interaction between two plates joined by a loose bolt. The bolt is assumed firmly attached to one of the plates with the opposite end free. Hence it resembles an arm that vibrates between two limiting surfaces. Both arms are fixed at one end and free to contact on the other end. Structure is excited harmonically at point A, see Figure 2.
3.2 Model II

This model is devised to explain the behavior of a bolted joint when the bolt holds the plates with reduced tightening force (see Figure 3). Different levels of tightening force were evaluated. Friction between components was also evaluated.

The same mesh as in Model I was used with the same boundary conditions and appropriate location of contact surfaces. Tightening was created in this way: a rotation was enforced in points C and D, located at a distance of 0.005 m from de T-junction on the fork’s upper and lower branches. Again, harmonic displacement was enforced in point A. The output of the system is measured in the same point as in Model I.

![Figure 3: Model II](image)

3.3 Model III

In this model, the interaction between plates of the joint was studied assuming bolts are tight but plates can hit each other in the space between them.

The structure has the shape of a box with two beams (Figure 4), with its ends fully fixed. A harmonic displacement is enforced in point E located at a distance of 0.005 m from right end. The output the system is measured at point F, located 0.03 m right from this structure’s left end.

The upper and lower parts of the box simulate the plates. Different material properties (stiffness and density) and cross sections were used for the upper and lower parts. This was necessary in order to make the parts come into contact during vibration.
4 RESULTS AND DISCUSSION

4.1 Model I

In this simplified model the interaction between plates and a loose bolt was studied. The time series of a typical output of the system due to a 50 Hz harmonic load on point A is depicted in Figure 5.

The frequency of the process in the figure is 50 Hz. The sinusoidal part of the waveform shows the contact between the loose plate and the bolt head or the fixed member. The amplitude of the sinusoidal displacement depends almost exclusively on the stiffness of the structure (our research showed that this is valid for excitation frequencies up to $1/10^{th}$ of the resonance frequency of the first mode of the structure). It is thus possible to estimate the amplitude of this displacement by performing a static test, in which an equivalent enforced displacement is applied to the structure.

The frequency of the sinusoids observed in Figure 5 is clearly higher than the excitation frequency since they have to complete a half period in less time than the half period of the
whole process \((1/50 \text{ Hz}/2 = 0.01 \text{ s})\) in the case in Figure 5.

The rationale for deriving a formula that estimates the angular frequency of the sinusoids is the following: the time that it takes for the sinusoid to complete its half period is given by half the period of the harmonic excitation load minus the time it takes for the tip of the loose member to cover the clearance between the bolt head and the fixed member.

The angular frequency of the sinusoids can be estimated with this formula:

\[
\omega_{\text{sinusoids}} = \frac{\pi}{\pi - \frac{1}{2} \arcsin \left( \frac{\text{Gap}}{A} \right)}
\]  

(1)

Where:

- \(\omega\) is the angular frequency of excitation
- \(\text{Gap}\) is half the total clearance between the upper and lower members of the joint (see Figure 2)
- \(A\) is the maximum displacement which the tip of the loose member would experience if the bolt was not present to limit the excursion, for a given enforced displacement at point A.

A similar analysis produces a formula for the phase of the sinusoids (the sinusoids start with a delay in the output signal):

\[
\phi_{\text{sinusoids}} = -\frac{\arcsin \left( \frac{\text{Gap}}{A} \right)}{\omega} \cdot \omega_{\text{sinusoids}}
\]  

(2)

The plateau areas in Figure 5, where oscillations are observed, occur when no contact exists between the plates. The observed oscillations are free vibrations of the fork structure.

The signal in Figure 5 is quasi-periodic so its spectrum is approximately discrete. An FFT was performed on the output. A portion of the spectrum is shown in Figure 6 for a 50 Hz excitation frequency. Only odd harmonics are present due to signal oddity.

![Figure 6: Spectrum Model I – Experimental vs. Theoretical Harmonics Amplitudes](image)
4.1.1 Analytical formulation of Model I

We propose an approximate formulation to predict the amplitude of the harmonics based on geometrical, physical and excitation parameters of the joint under study. We can approximate the shape of half a period of the output signal (as seen on Figure 5) and then expand this signal into a half-range Fourier series.

We model the signal analytically by making some assumptions: the plateau is considered free of oscillations and the frequency and phase of the sinusoidal part of the signal is reconstructed with equations (1) and (2). Enforcing a static displacement at point A, the amplitude of the sinusoid is obtained approximately.

A half period of this analytical piecewise function is defined as:

\[
f(t) = \begin{cases} 0, & 0 \leq t < \frac{\sin^{-1}(\text{Gap}_A)}{\omega} \\ \text{Amp} \cdot \sin(\omega \sinoids \cdot t + \phi_\sinoids), & \frac{\sin^{-1}(\text{Gap}_A)}{\omega} \leq t < \frac{\pi}{\omega} - \frac{\sin^{-1}(\text{Gap}_A)}{\omega} \\ 0, & \frac{\pi}{\omega} - \frac{\sin^{-1}(\text{Gap}_A)}{\omega} \leq t < \frac{\pi}{\omega} \end{cases}
\]

Where:

\( \text{Amp} \) is the amplitude of the sinusoid obtained by the static test.

By expanding this analytic function into a half-range Fourier series, we obtain the theoretical amplitude of the harmonics. The amplitude of the fundamental \((b_0)\) and \(n\)th harmonic \((b_n)\) are:

\[
b_0 = -\frac{8 \cdot \text{Amp} \cdot \text{Gap} \cdot \arccos\left(\frac{\text{Gap}}{A}\right)}{A \cdot \pi^2 + 4A \cdot \arccosh\left(\frac{\text{Gap}}{A}\right)^2} \quad (4)
\]

\[
b_n = -\frac{4 \cdot \text{Amp} \cdot \arccos\left(\frac{\text{Gap}}{A}\right) \cdot \left[ \sin\left(\pi \cdot \arcsin\left(\frac{\text{Gap}}{A}\right)\right) \right] + \sin\left(n \cdot \arcsin\left(\frac{\text{Gap}}{A}\right)\right)}{\pi^2 + 4n^2 \cdot \arccosh\left(\frac{\text{Gap}}{A}\right)^2} \quad (5)
\]

Equations (4) and (5) were used to compare theoretical vs. experimental amplitudes of harmonics in Model I. Geometrical and physical parameters were varied to check the accuracy of the proposed approach.

Table 1 shows the results. The table’s entries are percentage errors of theoretically obtained \(b_n\) relative to experimental ones. Figure 5 shows the amplitude of the experimental and theoretical harmonics for the case marked in red in Table 1.
Table 1: Model I Formula validation results

<table>
<thead>
<tr>
<th>Gap (m)</th>
<th>0.006</th>
<th>0.004</th>
<th>0.002</th>
<th>Constant Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fundamental</td>
<td>3%</td>
<td>2%</td>
<td>1%</td>
<td>E = 800 Gpa</td>
</tr>
<tr>
<td>3rd</td>
<td>-3%</td>
<td>-6%</td>
<td>-6%</td>
<td>Density = 3000 Kg/m3</td>
</tr>
<tr>
<td>5th</td>
<td>-42%</td>
<td>2%</td>
<td>-12%</td>
<td>Amplitude = 0.0025 m</td>
</tr>
<tr>
<td>E (Gpa)</td>
<td>200</td>
<td>500</td>
<td>800</td>
<td>Frequency = 50 Hz</td>
</tr>
<tr>
<td>Density (kg/m3)</td>
<td>3000</td>
<td>6000</td>
<td></td>
<td>Gap = 0.006 m</td>
</tr>
<tr>
<td>Fundamental</td>
<td>-1%</td>
<td>-3%</td>
<td>3%</td>
<td>E = 800 Gpa</td>
</tr>
<tr>
<td>3rd</td>
<td>-2%</td>
<td>-4%</td>
<td>-5%</td>
<td>Density = 3000 Kg/m3</td>
</tr>
<tr>
<td>5th</td>
<td>-7%</td>
<td>24%</td>
<td>58%</td>
<td>Amplitude = 0.003 m</td>
</tr>
<tr>
<td>Amplitude (m)</td>
<td>0.0025</td>
<td>0.003</td>
<td>0.0035</td>
<td>Frequency = 50 Hz</td>
</tr>
<tr>
<td>Frequency (Hz)</td>
<td>30</td>
<td>50</td>
<td>100</td>
<td>Gap = 0.006 m</td>
</tr>
<tr>
<td>Fundamental</td>
<td>-3%</td>
<td>-3%</td>
<td>-3%</td>
<td>E = 800 Gpa</td>
</tr>
<tr>
<td>3rd</td>
<td>-4%</td>
<td>-6%</td>
<td>-9%</td>
<td>Density = 6000 Kg/m3</td>
</tr>
<tr>
<td>5th</td>
<td>26%</td>
<td>19%</td>
<td>0%</td>
<td>Amplitude = 0.003 m</td>
</tr>
</tbody>
</table>

Where:

**Gap**: is half the total clearance between the bolt head and the fixed member (see Figure 2)

**E**: is Young’s modulus.

**Density**: is the density of the materials in the model.

**Amplitude**: is the amplitude of the displacement enforced on point A.

**Frequency**: is the frequency of excitation.

In Table I, only one parameter is changed at a time. The “Constant Parameters” box shows the values of these constant parameters for each case.

### 4.1.2 Analysis of results

Equations (4) and (5) approximate the amplitude of the fundamental and of the 3er harmonic precisely. The 5th harmonic showed some differences, especially for high stiffness values. This can be attributed to the idealization of a flat plateau. The hitting process of the loose plate with either end introduces high frequency oscillations in the time domain, effect not considered in the theoretical solution.
4.2 Model II

In this model, the behavior of a bolted joint where the plates are held firmly by the bolt was studied. Different levels of tightening force were analyzed. The same mesh as in Model I was used. The material properties are stated in Table 2. In order to hold the loose member firmly, the upper and lower branches of the fork structure were rotated. Once the branches were closed, a harmonic displacement was enforced at point A. When the loose member was firmly held, both parts oscillated in unison. When the closing force was low, loss of contact occurred between parts.

<table>
<thead>
<tr>
<th>Width square cross section (m)</th>
<th>Young Modulus (GPa)</th>
<th>Density (kg/m³)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.005</td>
<td>800</td>
<td>3000</td>
</tr>
</tbody>
</table>

4.2.1 Harmonic Distortion

There is a relation between bolt tightening force on a bolted joint and the appearance harmonic distortion. To quantify the amount of harmonic distortion, an index named Harmonicity Index ($C_{\text{harmonic}}$) is introduced. This index is the quotient between the energy in the FFT’s bins corresponding to harmonics of the excitation frequency and the energy in all the other bins (up to a certain frequency). The fundamental – which oscillates at the frequency of excitation – is not taken into account in the energy of the harmonics because it is present in all the spectrums. The FFTs were performed on the output.

The mean contact force on the node located on the tip of the upper branch was used to approximate the tightening force during the simulation.

Increasing tightening forces were applied. Two excitation frequencies, 50 and 100 Hz, were enforced. The results are summarized in Figure 7.

![Figure 7: Harmonicity index vs. Contact Force (tip node)](image)

Increasing the tightening force reduces the Harmonicity Index as expected. Loss of contact
was present only in the two lowest values of mean contact force shown in Figure 7.

4.2.2 Use of a Linear Structure

We analyze the effect which an ideally joined structure would have on the obtained results. The proposed linear structure has a common point at the position where the upper and lower branches get in contact when closed. A harmonic displacement is enforced on the same point as before. No harmonic distortion appears in the analysis of this structure since it is linear.

We define Amplitude Transmissibility as the ratio of the rms amplitude at the output of the contacting model, relative to the rms output of the linear structure. Unit value for the amplitude transmissibility indicates that the joint transmits the same amplitude of vibration at its output as the linear structure does.

Figure 8 shows the results of the Amplitude Transmissibility for increasing tightening forces. A coefficient of friction of 0.4 between contacting parts was also introduced.

![Amplitude Transmissibility Index vs. Mean Contact Force](image)

Figure 8: Amplitude Transmissibility vs. Contact Force (tip node)

When replacing the joint with a linear connection, Amplitude Transmissibility could be underestimated. We believe that Amplitude Transmissibility grows with increasing tightening due to reduced stiffness of the structure at low tightening values.

Friction dissipates energy and since we are exciting the model with an enforced displacement, high values of tightening force dissipate more energy due to high friction forces, reducing displacements. This is seen as a reduction of the Amplitude Transmissibility.

There is no appreciable change in the results by changing the frequency of excitation.

4.3 Model III

This model explains the behavior of a bolted joint in which the plates are loose and can hit each other.

Material properties were selected in order to make contact possible between the upper and lower plates. Properties are listed in Table 3.
Table 3: Material Properties and Cross Sections for Model III

<table>
<thead>
<tr>
<th>Plate</th>
<th>Width cross section (square) (m)</th>
<th>Young Modulus (GPa)</th>
<th>Density (kg/m³)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upper</td>
<td>0.001</td>
<td>100</td>
<td>13000</td>
</tr>
<tr>
<td>Lower</td>
<td>0.002</td>
<td>200</td>
<td>10000</td>
</tr>
</tbody>
</table>

Three submodels where developed, with different geometries. The aspect ratio of the box - depicted in Figure 4 - was devised so that plate could hit each other with increasing energy in each case (see Table 4). Also listed is the effective gap between the plates. Output was measured at point F. Three different excitation frequencies were used: 150, 200 and 300 Hz. Analysis was performed with and without contact to compare impact effects.

Table 4: Aspect ratios

<table>
<thead>
<tr>
<th>Submodel</th>
<th>Aspect Ratio “box” (Length/Width)</th>
<th>Gap (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10</td>
<td>0.001</td>
</tr>
<tr>
<td>B</td>
<td>12.5</td>
<td>0.001</td>
</tr>
<tr>
<td>C</td>
<td>25</td>
<td>0.0006</td>
</tr>
</tbody>
</table>

4.3.1 Analysis of results

Figure 9 shows the spectrum of submodels “A”, “B” and “C” with and without contact (excitation frequency not shown). A few observations can be made; further research is being conducted for better characterization:

- In general, the effect of harmonically driving a joint whose loose members can contact each other is to excite the modal behavior of the structure.
- Introducing a contact condition between the members has similar effect to the presence of damping: it spreads the energy around the resonant frequency.
- Resonant peaks shift frequency and reduce their amplitude when contact is introduced.

Figure 8: Spectrum - submodels A, B and C (with and without contact)
5 CONCLUSIONS

Three models were presented towards an understanding of the dynamic behavior of bolted joints with nonlinearities. These models help understand the phenomena present in a joint when dynamically excited.

Model I simulated the interaction between the plates in a joint with a loose bolt. Harmonic distortion appeared due to the nonlinear contact condition. This indicates that in order to model a real joint, harmonics of the excitation frequency must also be taken into account. A heuristic formulation was developed for estimating the amplitude of the harmonics.

Model II analyzed a joint model in which the plates are held firmly by the bolt. Increasing the tightening force lowers the amount of harmonic distortion, as described by the Harmonicity Index. A linear structure was proposed as a substitute for the contact joint. The use of this structure underestimates the amplitudes of vibration of the real contacting joint.

Model III studied the interaction between loose plates. The overall outcome of driving this structure with a harmonic displacement is to excite the modal behavior of the structure. The effect of the contact condition was to spread energy around the natural frequencies and to lower the amplitudes of the resonance. A shift in frequency of the resonant peaks was also observed.

Future work will focus on improving the heuristic formulation and developing a user-defined element which incorporates many of the characteristics analyzed in the present work.

REFERENCES

bolted joints under harmonic loading”. Computers and Structures 84 (2005) 25–33
ON THE NONLINEAR DYNAMICS OF NETWORK STRUCTURES UNDER STOCHASTIC EXCITATION

Wagner de Cerqueira Leite, Reyolando M.L.R.F. Brasil

¹ Universidade Federal do ABC - UFABC
R. Arcturus, SN, Bloco delta, Sala 386 CEP 09606-070, S. Bernardo do Campo, SP
wagner.leite@ufabc.edu.br

² Universidade Federal do ABC - UFABC
R. Arcturus, SN, Bloco delta, Sala 386 CEP 09606-070, S. Bernardo do Campo, SP
reyolando.brasil@ufabc.edu.br

Key words: cables structures, nonlinear dynamics, random vibrations.

Abstract. In this paper, the dynamic analysis of a cable structure under wind action is made. Given the magnitude of the displacements, the behavior of this structure is geometrically nonlinear. The wind forces are divided into two parts, one average parcel in time and another parcel of random nature. This wind forces randomness is simulated through a process called "synthetic wind." The equations of motion, in the vector treatment used, are resolved through a numerical method, the Central Difference Method. As a final result, displacements series are obtained to the nodes of the structure and normal tensions series to the cables, allowing the determination of characteristic values to be employed in the analyzed structure design.

1 INTRODUCTION

The cable structure design is of great interest because it combines low weight with high rigidity to forces that induce tension in its elements. Among the numerous examples of its application, we have the cable networks like the one studied here, cable-stayed bridges, and tensegrity structures. Often, as with the example of this paper, are used to cover large spans and submitted to great intensity of loads, resulting displacements of such magnitude that the consideration of geometric nonlinearity is mandatory. For the vector formulation used herein, the equations of motion are solved by the central difference method for successive instants (calculation steps), making up the update of geometry of the structure at each instant. The determination of the initial shape of the structure, under permanent loading, is a calculation that is beyond the trivial to most the structural engineers. A possible solution to this problem is the use of dynamic relaxation, which is adopted here.

2 EQUATIONS OF MOTION

The ends of cable elements are considered the nodes $n_i$ of the structure, $n_{i_1} i = 1 \text{ to } NN$, where $NN$ is the number of nodes of the structure. The dynamic loads are applied at these
nodes as concentrated forces with $P_{i,1}(t), P_{i,2}(t), P_{i,3}(t)$ components referred to a right-handed axes coordinate system $e_1, e_2, e_3$, and the masses and the damping as nodal lumped masses $M_i$ and dampers $C_i$.

Knowing the coordinates $Y_{i,1}, Y_{i,2}, Y_{i,3}$ for each node $n_i$ on the system $e_1, e_2, e_3$, the lengths of the cable elements are defined from the coordinates of its initial and final nodes. For a cable element $j$, $j = 1$ to $NE$, where $NE$ is number of cable elements, with initial and final nodes $NI$ e $NF$, the length $L_j$ is given by Eq. (1):

$$L_j = \sqrt{(Y_{NF,1} - Y_{NL,1})^2 + (Y_{NF,2} - Y_{NL,2})^2 + (Y_{NF,3} - Y_{NL,3})^2}$$ (1)

Figure 1: The cable element

At figure 1, $L_j$, $A_j$ and $E_j$ are, respectively, length, sectional area and the Young’s modulus of the material of a cable $j$.

With the nodes coordinates, the direction cosines $\cos \alpha_{j1}, \cos \alpha_{j2}, \cos \alpha_{j3}$ of the cable axis, related to the axes system $e_1, e_2, e_3$, are determined.

The stress-strain ($\sigma \times \varepsilon$) diagram adopted is shown at figure 2, where $\sigma_E$ is the yield stress:

Figure 2: Stress-strain relationship

Defined the initial geometry of the structure to be analyzed, the displacements of the structural nodes are obtained at successive instants (steps). In this manner, series of displacements will be generated. For this, the Eq. (2) of motion is solved for each node at each time step.

$$M \ddot{x} + C \dot{x} + K x = P(t)$$ (2)
This equation is written for the nodes \( n_i \) at the three orthogonal directions in the space, \( e_1, e_2, e_3 \).

Let \( x_{i,1}(t), x_{i,2}(t), x_{i,3}(t) \) be the displacement components for a node \( n_i \), referred on axes \( e_1, e_2, e_3 \). Thus, \( x_{i,1}'(t), x_{i,2}'(t), x_{i,3}'(t) \) and \( x_{i,1}''(t), x_{i,2}''(t), x_{i,3}''(t) \), where upper lines indicate derivation in time, are velocity and acceleration components, \( M_i, x_{i,1}'(t), M_i, x_{i,2}'(t), M_i, x_{i,3}'(t) \) are the nodal inertial forces, and \( C_i, x_{i,1}'(t), C_i, x_{i,2}'(t), C_i, x_{i,3}'(t) \) are the nodal damping forces.

A numerical method to solve these equations is the Central Difference Method, as presented by Brasil [2]. Let us consider the two expansions of Eq. (4) in Taylor’ series at the neighborhood of a nodal displacement \( x_{ik,t} \), where \( x_{ik,t} \) is a displacement component on a referential axis direction \( k, k = 1 \) to \( 3 \), and in a given instant \( t \), being \( x_{ik,t} \) supposed known:

\[
x_{ik,t+\Delta t} = x_{ik,t} + \Delta t x_{ik,t}' + 0.5 \Delta t^2 x_{ik,t}'' + \ldots \quad (3)
\]

and the sum of expansions (4) results an approximation to accelerations:

\[
x_{ik,t}'' = (x_{ik,t+\Delta t} - 2 x_{ik,t} + x_{ik,t-\Delta t})/\Delta t^2 \quad (5)
\]

Rewriting the generic equilibrium equation (2) for a node \( n_i \) in \( k \) direction and time \( t \), where \( F_{ik,t} \) is the sum of force components in \( k \) direction from cables hinged at node \( n_i \):

\[
M_i, x_{ik,t}'' + C_i, x_{ik,t} + F_{ik,t} = p_{ik,t} \quad (6)
\]

and replacing Eq. (4) and Eq. (5) in equation (6), one arrives at the pseudo static equilibrium equation:

\[
x_{ik,t+\Delta t} = [p_{ik,t} - F_{ik,t} + (2, M_i/\Delta t^2) x_{ik,t} - (M_i/\Delta t^2 - C_i/2 \Delta t) x_{ik,t-\Delta t}]/(M_i/\Delta t^2 + C_i/2 \Delta t) \quad (7)
\]

Equation (7) will be used to determine the displacements of the nodes at successive time steps that corresponds to successive time increments \( \Delta t \). Note that because the calculation process depends only on values of displacements available, obtained in previous steps, is an explicit method. But as it requires \( x_{ik,t-\Delta t} \), it is not self-initiated.

### 2.1 Start of the Calculation Process

To start the process, acceleration at the initial time \( x_{ik,0}'' \) is obtained from the equation of motion and initial conditions \( x_{ik,0}, x_{ik,0}', x_{ik,0}'' \). Then, by supposing constant acceleration \( (x_{ik,t=0+\Delta t} = x_{ik,t=0}'') \) in the first interval, it results:

\[
x_{ik,t=0+\Delta t} = x_{ik,t=0} + \Delta t x_{ik,t=0}'
\]

\[
x_{ik,t=0+\Delta t} = x_{ik,t=0} \Delta t^2/2 + x_{ik,t=0}' \Delta t + x_{ik,t=0}'' \quad (8)
\]
2.2 Updating of geometry and forces $F_{ik,t+\Delta t}$ to the next step

With the displacements thus obtained, the new coordinates $y_{ik,t+\Delta t}$ for the nodes are calculated in a given time and, therefore, the new lengths $L_{ij,t+\Delta t}$ of the elements are determined with Eq. 1, from which one obtains the length variations $\Delta L_{ij,t+\Delta t}$ of the cables.

$$\Delta L_{ij,t+\Delta t} = L_{ij,t+\Delta t} - L_{ij,t}$$  \hspace{1cm} (9)

Also new direction cosines for the axes of the elements $cos \alpha_{jk,t+\Delta t}$ are calculated.

It is admitted that normal forces in compressed cables are negligible. Considering this, cables “in compression” at one given step of calculation must not be considered in the structure, making its normal forces equal to zero during the step. To detect a cable “in compression”, its length, the distance between cable ends, is compared with the length of this cable when its normal force is equal to zero (untensioned length $L_{jN0}$).

The following algorithm gives an estimate of the normal force at the time $t + \Delta t$:

If the $j$ element is a cable, and $L_{jN,t+\Delta t}$ is less than or equal to the cable unstressed length $L_{jN0}$, then the normal force in the cable sections, $N_{j,t+\Delta t}$, is zero.

otherwise:

$$\vec{N}_{j,t+\Delta t} = (E \cdot A_j \cdot \Delta L_{j,t+\Delta t} / L_{j,t}) + N_{j,t}$$

and then:

$$If \quad \vec{N}_{j,t+\Delta t} > N_{Ej}, \text{where} \ N_{Ej} = \sigma_E \cdot A_j, \text{so} \ N_{j,t+\Delta t} = N_{Ej}$$

Or else $N_{j,t+\Delta t} = \vec{N}_{j,t+\Delta t}$  \hspace{1cm} (10)

The program developed accepts bar elements in the structure. Normal compression forces in bar elements, when it occurs, may be significant. Thus, if an element is not a cable, but a bar, the following condition must be added to the algorithm:

$$If \ N_{j,t+\Delta t} < -N_{Ej}, \text{so} \ N_{j,t+\Delta t} = -N_{Ej} \text{ (only bar element)}$$

If the $j$ element is a cable element and $N_{j,t+\Delta t} = N_{Ej}$, then the length $L_{jN0}$ is updated for the next iterations:

$$L_{jN0} = L_{j,t+\Delta t} / \left(1 + N_{Ej} / E_j \cdot A_j \right)$$  \hspace{1cm} (11)

Projecting the normal forces $N_{j,t+\Delta t}$ applied to node $n_i$ in the direction $k$ through its direction cosines $cos \alpha_{jk,t+\Delta t}$, we obtain $F_{ik,t+\Delta t}$ and proceed to a new iteration, writing the equilibrium equation for time $t + \Delta t$.

3 STATIC DISPLACEMENTS. DYNAMIC RELAXATION.

The following procedure is based on an analogy between static and dynamic structural analysis.

The static analysis of structures of nonlinear behavior in general is iterative. From the nonlinear equations of nodal equilibrium, successive calculations of the displacements of the nodes are made until its convergence to the equilibrium position.

Here, in a dynamic analysis, the structure, considered to be critically damped, is submitted to an instant loading until displacements converge to the end position of static equilibrium.
This convergence is checked by observing the time after which increasing of the displacement becomes very small and can be neglected.

To determine the critical damping of the structure, we need to determine its circular natural frequency of vibration $\omega$. Hence, the critical nodal damping $C_{ci}$ on a node $n_i$ is:

$$C_{ci} = 2 M_i \omega$$

(12)

An estimate of the natural frequency may be made by an analysis of the structure in time domain through the application of a load from a given instant whose intensity is held constant over a time interval, being this load removed in an instant later and then allowing the free vibration of the structure, while the nodal damping is made equal to zero.

The resulting time history is then typical of a periodic function with constant amplitude. Plotting this curve, one can estimate the period $T$ through measurement made directly on the curve.

The cyclic frequency, $f$, is

$$f = 1/T$$

(13)

and the circular frequency is

$$\omega = 2 \pi f$$

(14)

Substituting in Eq. (14), one gets the critical damping $C_c$.

Next, displacements in time for the structure are computed, with estimated nodal critical damping and permanent loads. Plotting this time history it is possible to verify if the critical damping value is correct.

If the estimated damping value is slightly below the real critical damping (underdamped structure), displacements will oscillate, tending to the equilibrium position. If the value of the damping is above the real (overdamped structure), convergence occurs without exceeding the position of static equilibrium. In both cases, however, the convergence will occur on a longer time than that for critical damping.

4 CABLE ROOF SUBMITTED TO TURBULENT WIND

Here, we apply the central difference method to the analysis of a hyperbolic paraboloid shaped cable network which constitutes the structure of a roof, subjected to wind forces. Figure 4 corresponds to that presented by Paik [6], where the lengths are in feet, as in the original reference, also presented by Esquilland and Saillard [4]. Table 1 and its results to dead load displacements for this structure are those presented originally by Paik [6].
Table 1 - Dead load displacements according Paik [6]

<table>
<thead>
<tr>
<th>Nodes</th>
<th>Initial coordinates</th>
<th>Dead load displacements</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(x)</td>
<td>(y)</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>15</td>
</tr>
<tr>
<td>7</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>15</td>
<td>20</td>
</tr>
</tbody>
</table>

Self weight + covering: 5 psf = 240 N/m²
Prestressing on x direction: 60 k = 2.67 \times 10^5 N
Prestressing on y direction: 60 k = 2.6713 \times 10^5 N
Cables section area: 1.4 in² = 9.03 \times 10^{-4} m²
Young modulus of cable material: 24000 ksi = 1.657 \times 10^{11} N/m²
We consider that to each node, from 1 to 25, it is applied the weight of an area whose horizontal plane projection is a 40 ft (12.192 m) sides square. The paraboloid surface area is given by:

\[ A = \int \int [1 + (dz/dx)^2 + (dy/dx)^2]^{1/2} \, dx \, dy \]  

(17)

That for the geometry given by Eq. (18) results:

\[ A = \int \int [1 + (x/390)^2 + (y/390)^2]^{1/2} \, dx \, dy \]  

(18)

The term inside the integration may be simplified, if considered as a function of two variables developed in a Taylor series, \( f(u, v) = (1 + u^2 + v^2)^{1/2} \). Considering terms of the series to the derivative of order 2, one obtains:

\[ A = \int \int [(1 + (x/390)^2 + (y/390)^2)]^{1/2} / 2 \, dx \, dy \]  

(19)

Finally, by integrating between the limits \( x_1 \) and \( x_2 \), and \( y_1 \) and \( y_2 \) one has the contribution area to a node:

\[ A = (x_2 - x_1) \, (y_2 - y_1) \{1 + [(x_2 - x_1)^2 + (y_2 - y_1)^2]/912,600\} = \\
= 12.192 \, 12.192 \, [1 + (12.0192^2 + 12.0192^2)]/912,600 = 148.69 m^2 \]

Thus, on each node, self-weight is equal to the contribution area multiplied by the weight per unit area, 148 69 240 = 35,615 N, and the nodal mass, equal to the weight divided by the acceleration of gravity, is 3,632 kg.

### 4.2 Wind forces

We determine wind forces from the wind pressures, that are proportional to the square of the wind velocities modified by drag coefficients \( C_p \). These coefficients \( C_p \) are the result of testing a model of the structure in a wind tunnel. We used the coefficients \( C_p \) given by Esquilland and Saillard [4] and that was used, too, by Paik [6].

According the NBR6123 Brazilian standard “Forças Devidas Ao Vento Nas Edificações” (Wind Forces on Buildings), for a \( V_k \) characteristic wind speed, we get a \( q_k \) characteristic wind pressure given by:

\[ q_k = 0.613 \, V_k^2 \]  

(20)

Thus, the pressure over a structural point is given by:

\[ p = C_p \, q_k = 0.613 \, C_p \, V_k^2 \]  

(21)

The pressure \( p \) on a structural node will be multiplied by its contributing area to result the wind force on the node. The wind velocity, and so its pressure and force, is composed of two parcels, one being a time average velocity or pressure, and the other is a floating contribution about the average (wind gust).

It will be considered that the average parcel renders a permanent load, which will be applied to the structure with its self-weight to determine the initial geometry, while the floating parcel gives the dynamic force.

The structure analyzed in this paper will be submitted to a wind force of maximum intensity. This force corresponds to a peak wind velocity, or pressure peak, which is believed
to occur at a given instant. The design standards, however, only provide average wind velocities in time intervals. Thus, it was decided to adopt as a peak velocity the average for the shortest time interval of NBR6123 standard, which is 3 seconds, $V_3$. This peak speed $V_3$ includes the mentioned average and the floating parcels (gusts).

It will be assumed that the peak average velocity is the standard average for a much longer time interval, adopted equal to 10 minutes or 600s, which will be denoted $V_{600}$.

Having the ratio between the average and the peak velocities, one has also the ratio between the floating and the peak parcels. The NBR6123 standard provides, for flat terrain, group II construction, category IV terrain, a ratio between the velocities $V_{600}$ and $V_3$ equals to: \[ V_{600}/V_3 = 0.60 \] (22)

from where:

\[ P_{600}/P_3 = (V_{600}/V_3)^2 = 0.36 \] (23)

In the same conditions, $V_3$, the peak velocity, on a point at $z$ meters height above the ground is, according to NBR6123:

\[ V_3 = 33.25 \frac{z}{10} 0.12 \text{ m/s} \] (24)

Then, the peak pressure is:

\[ p = 0.613 \ C_p \ V_3^2 = 637.56 \ C_p \ \frac{z}{10} 0.24 N/m^2 \] (25)

and, therefore, the average parcel of the peak pressure is:

\[ \bar{p} = 0.36 \ p = 229.52 \ C_p \ \frac{z}{10} 0.24 N/m^2 \] (26)

and the floating parcel is:

\[ p' = 0.64 \ p = 408.04 \ C_p \ \frac{z}{10} 0.24 N/m^2 \] (27)

To have the correspondent forces on the nodes, it is necessary to multiply these pressures by the node contribution area $A = 148.69 \text{ m}^2$, resulting to the average parcel of wind:

\[ \bar{P} = 3.41 \times 10^4 \ C_p \ \frac{z}{10} 0.24 N/m^2 \] (28)

and the floating parcel or gust:

\[ P' = 6.10 \times 10^4 \ C_p \ \frac{z}{10} 0.24 N/m^2 \] (29)

The drag coefficients $C_p$ for this structure, obtained for different wind directions from wind tunnel tests, are given by Esquilland and Saillard [4], and used by Paik [6]. It will be considered in this paper, for simplicity, the coefficients in just one of these directions.

**Table 2 - Drag coefficients according Esquilland and Saillard, 1963**

<table>
<thead>
<tr>
<th>Node</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>0</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_p$</td>
<td>-.50</td>
<td>-.33</td>
<td>-.45</td>
<td>-.60</td>
<td>.20</td>
<td>.10</td>
<td>-.36</td>
<td>-.68</td>
<td>-.83</td>
<td>.10</td>
<td>.43</td>
<td>.44</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Node</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_p$</td>
<td>.10</td>
<td>-.33</td>
<td>-.63</td>
<td>-.80</td>
<td>-.83</td>
<td>-.68</td>
<td>-.36</td>
<td>-.10</td>
<td>.20</td>
<td>-.60</td>
<td>-.45</td>
<td>-.33</td>
<td>-.50</td>
</tr>
</tbody>
</table>
4.3 Random Nature of Wind Simulation

The Power Spectral Density function adopted is the reduced spectrum of the National Building Code of Canada, which is a slightly modified power spectrum of Davenport to the wind speed.

\[ n \frac{S_p(n)}{u^2} = 4 \frac{x}{(1 + x^2)^{4/3}}; \quad x = 1220 \frac{n}{\bar{V}(10)} \quad (30) \]

\(\bar{V}(10)\) is the average wind speed at a height \(z = 10\) m in the open.

This Power Spectral Density function of wind velocity can be used to determine the amplitudes of the harmonics instead of the spectral density function of the wind pressure on a point of the structure, given that the two functions differ only by a constant, in which it is built in the drag coefficient of the structure to this point.

Considering that the floating parcel of peak pressure is equals to a sample function constituted of an infinite number of harmonic functions, it is possible to represent this variable parcel through the Fourier integral:

\[ p'(t) = \int_{-\infty}^{+\infty} C \cos(2\pi n t - \theta(n))dn \quad (31) \]

where:

\[ C(n) = [A^2(n) + B^2(n)]^{1/2} \]
\[ \theta(n) = tg^{-1}(B(n)/A(n)) \]
\[ A(n) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} p'(t) \cos(2\pi n t)dn \]
\[ B(n) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} p'(t) \sin(2\pi n t)dn \quad (32) \]

The variance or mean square deviation of \(p'(t)\), calculated over a sufficiently large time interval \(t\), is equal to:

\[ \sigma^2(p') = \frac{1}{T} \int_{-T/2}^{+T/2} p'(t)^2dt = \frac{2}{T} \int_{0}^{\infty} C^2(n)dn \quad (33) \]

Making \(T\) tend to infinity, one can write:

\[ \sigma^2(p') = \int_{-\infty}^{+\infty} S_p(n)dn \quad (34) \]

where \(S_p(n)\) is the spectral density function of \(p'(t)\), with \(S_p(n)\) representing the contribution of a frequency interval \(dn\) for the average quadratic deviation \(\sigma^2(p')\).

Instead of an infinite number of functions that can reproduce it perfectly, in this process \(p'(t)\) is represented by a \(k\) finite number of harmonic functions with periods covering the range of duration of gusts, from 0.5s to 600s. According Franco [5] one of these functions, the \(rth\), must have the resonance period \(T\) of the structure. The other functions periods must be multiples or submultiples of \(T\) by powers of two, so that in a logarithmic scale, it results bands of equal length of spectrum for each function, as shown Fig 7.
The equation, then, becomes:

\[ p'(t) \cong \sum_{k=1}^{m} C_k \cos \left( \frac{2\pi}{T_e T_k} t - \theta_k \right) \tag{35} \]

where \( m \) is the number of harmonic functions and bands of the spectrum, and:

\[ C_k = \sqrt{\frac{2}{(k)}} \int S(n)dn \tag{36} \]

\[ r_k = 2^{k-r} \tag{37} \]

The \( C_k \) value is calculated by integrating numerically within the band (interval) of the spectrum corresponding to it.

Taking \( p' \) as the maximum amplitude of the variable parcel of the pressure, given by Eq.27, to each one of its harmonic components in the sample function corresponds a maximum pressure amplitude:

\[ p_k = \frac{C_k}{m} p' = c_k p' \tag{38} \]

Values for \( \theta_k \) phase angles for each of the harmonic functions that constitute the sample function are taken from a series of randomly generated numbers. Thus, the harmonic functions that comprise variable wind pressure will overlap randomly, in accordance with the phase angle combinations.

This combination of harmonic functions is a simulation of a possible actual sample originated from wind measurements. Such a combination of harmonic functions is called "synthetic wind."

In this manner, the synthetic wind is given by:

\[ p'(t) \cong \sum_{k=1}^{m} p_k' \cos \left( \frac{2\pi}{T_e T_k} t - \theta_k \right) \tag{39} \]

Changing the \( \theta_k \) phase angles by values of another set of random numbers in a harmonic functions combination, we get another synthetic wind, simulating another sample function. In this paper, 20 sets of \( k=12 \) generated random number (\( \theta_k \) phase angles) were adopted, and so 20 different computer analysis were performed, each one to one different synthetic wind, resulting wind displacements and normal forces in the cables of the structure.
These synthetic winds must be applied to the most unfavorable point of the structure, where they will induce the largest displacements or stresses. This point is named gust center.

The program implemented in this study allows not only the generation of internal forces and displacements time histories for each sample function (synthetic wind), but also records the maximum values resulting from action of a given sample function (a given combination of phase angles in harmonic functions).

The maximum stresses and displacements obtained from the application of different synthetic winds are then adjusted to a maximums probability distribution, the Gumbel distribution:

$$f(x) = e^{-e^{(x-\mu)/\beta}}$$

(40)

where:

$$\beta = 6^{1/2} (\sigma/\pi)$$

and:

$$\alpha = \mu - 0.4500 \sigma$$

\(\mu\) is the mean, and \(\sigma\) the standard deviation.

Table 3 shows the maximum displacements determined for node 13, and the maximum tensions for the cable element (20) from the processing:

<table>
<thead>
<tr>
<th>Comb.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Displ. (m)</td>
<td>-1.814</td>
<td>-1.842</td>
<td>-1.838</td>
<td>-1.837</td>
<td>-1.811</td>
<td>-1.828</td>
<td>-1.855</td>
<td>-1.834</td>
<td>-1.801</td>
<td>-1.820</td>
<td>-1.849</td>
<td>-1.862</td>
</tr>
<tr>
<td>Tension (N)</td>
<td>700,613</td>
<td>699,184</td>
<td>706,582</td>
<td>693,841</td>
<td>695,526</td>
<td>700,573</td>
<td>694,917</td>
<td>696,618</td>
<td>698,828</td>
<td>702,190</td>
<td>698,973</td>
<td>698,690</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Comb.</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Displ. (m)</td>
<td>-1.829</td>
<td>-1.831</td>
<td>-1.802</td>
<td>-1.875</td>
<td>-1.857</td>
<td>-1.814</td>
<td>-1.808</td>
<td>-1.803</td>
</tr>
<tr>
<td>Tension (N)</td>
<td>701058</td>
<td>696462</td>
<td>706482</td>
<td>699480</td>
<td>695523</td>
<td>690177</td>
<td>702254</td>
<td>698554</td>
</tr>
</tbody>
</table>

Hence, for the displacements:

$$f(x) = e^{-e^{(x-1.823)/0.01683}}$$

(41)

and the characteristic value for design is the one with 5% probability of be exceeded, that is \(x = 1.773\)m.

For tensions:

$$f(x) = e^{-e^{(x-697198)/3013.89}} \rightarrow x = 688,246 \text{ N}$$

(42)

and among the combinations, the one whose displacement is the closest to the characteristic value of the displacements is combination 15. Regarding tensions, the combination is the 19.
For this reason, it is said that the characteristic wind for the displacement is the synthetic wind that corresponds to combination 12.
For tensions in the element 20, the characteristic wind corresponds to combination 19.

5 CONCLUSIONS

A time domain analysis of a cable structure presenting geometric nonlinearity was displayed using the central difference method. One way to consider the random nature of wind forces was proposed. As a final result, series of displacements and forces on the elements were generated, and from that, characteristic values for structural design were obtained. It is expected that, with the increasing use of cable structures, this paper can be a contribution for those who choose this structural solution.

REFERENCES

SIZE-DEPENDENT DYNAMIC ANALYSIS OF TIMOSHENKO MICROBEAM USING THE FINITE ELEMENT METHOD

DANIEL H. FELIX¹, GRACIELA I. GUERRERO² AND DIANA V. BAMBILL³

¹ Departamento de Ingeniería, Instituto de Mecánica Aplicada, Universidad Nacional del Sur
Avda. Alem 1253, 8000 Bahía Blanca, Argentina
dhfelix@uns.edu.ar, http://www.uns.edu.ar

² Facultad de Ingeniería, Universidad Nacional de la Patagonia San Juan Bosco
Km 4, 9005, Comodoro Rivadavia, Argentina
graisague@hotmail.es; http://www.unp.edu.ar

³ Departamento de Ingeniería, Instituto de Mecánica Aplicada, Universidad Nacional del Sur
Consejo Nacional de Investigaciones Científicas y Técnicas
Avda. Alem 1253, 8000 Bahía Blanca, Argentina
dbambill@criba.edu.ar, http://www.uns.edu.ar

Key words: Micro Structure, Couple Stress, Timoshenko Beam, Finite Elements.

Abstract. Micro beams have been widely used in the last decades, as components of MEMS (Micro Electro Mechanical Systems) and NEMS (Nano Electro Mechanical Systems). In the present work, a finite element formulation is developed for Timoshenko micro beams, taking into account the size-dependent effects. For this purpose the modified couple stress theory is used [1]. The modified couple stress theory takes advantage in simplicity, since it reduces the additional elasticity constants to only one [2], it is the length scale parameter. A dynamic study is carried out, considering the material length scale effects. The results are compared with those obtained by the classical elasticity theory. The first natural frequencies of beams with different boundary conditions are obtained and compared with available results in the technical literature.

1 INTRODUCTION

The dimensions of some components used in micro/nano electromechanical systems (MEMS/NEMS) are in the order of micrometers, or even of nanometers. When the material of such components is an elastic solid, it is named micro-elastic solid. The micro-elastic solids cannot be analyzed with enough accuracy, using the classic elastic theory (CET) [3], due the presence of complex forces and couples at micro scale level. So, in this case, a generalized elastic theory is needed that takes into account the size dependent material properties.

The generalized elastic theory is part of the generalized continuum theory, which was formally presented, for first time, in 1909 by the Cosserat brothers [4]. However this new theory began generating a great interest among engineers and researchers, starting from the 1960s, when many publications with different size-dependent theories, began appearing around the world.
In the sixties among these new theories, many couple stress theories \[3,5,6\] appear too. They consider in addition the presence of couples over each point of the continuum. Toupin said \[7,8\]: “In conventional mechanics, a force drives a material particle to translate, while in the couple stress theories for linear elastic materials, also a couple drives a material particle to rotate”.

These couple stress new theories are a little complicated to apply because of several additional parameters, which are considered to take into account the size-dependent material properties \[9\].

In 2002, the modified couple stress theory, (MCST) is presented by Yang et al \[1\], which introduced an additional equilibrium equation to take into account the couples’ behavior of \[10\].

This new size-dependent theory is applicable to linear elastic models for isotropic materials and it has a simpler implementation, since it has just one additional material constant \[2\], named the length scale parameter. A goal of this theory is that, it is in very good agreement with the experimental results reported in the literature \[11\].

In 2006 Park and Gao proposed MCST for the bending of an Euler-Bernoulli beam model, to capture the size effect \[12,13\] and in 2009 Kong et al \[14\] developed the modified strain gradient elasticity theory, based on Euler-Bernoulli beam model. In 2014 Guerrero et al, \[15\] analyzed with MCST the dynamic of an Euler-Bernoulli micro beam by the Ritz method and Felix et al analyzed an Euler-Bernoulli micro beam with non local elasticity theory, using the finite element method \[16\].

In the present work a numerical investigation of the size effects on free vibration of Timoshenko beams, based on the MCST, is carried out. For this purpose, the finite element method has been used, implementing a recently developed element, obtained by Kahrobaian et al \[17\]. the first two natural frequencies of Timoshenko micro-beams with different combinations of classical boundary conditions are calculated. The results are presented in non dimensional graphics and in a dimensional form, for the particular case of an epoxy material.

## 2 PROPERTIES OF A TIMOSHENKO MICROBEAM

### 2.1 Kinematic properties for the CET

The Figure 1 shows a scheme of the kinematic of a Timoshenko beam:

In the Timoshenko beam theory we take into account the shear deformation. In consequence, the distortion \( \gamma_{xx} \) of the cross section appears.

As it is shown in the above figure, the distortion \( \gamma_{xx} \), the rotation of the cross section \( \theta \) and the slope of the beam \( \partial w/\partial x \), are related by:

\[
\frac{\partial w}{\partial x} = \phi + \gamma_{xx} \tag{1}
\]

Two independent generalized displacements are considered in the Timoshenko beam model. They are the transversal displacement \( w \) and the rotation of the cross section \( \phi \). As function of them the axial displacement \( u \) can be expressed in the form:

\[
u = -z \phi \tag{2}
\]
Then, the non-zero components of the strain tensor can be expressed as follows:

$$
\varepsilon_{xx} = \frac{\partial u}{\partial x} = -z \frac{\partial \phi}{\partial x}
$$

(3)

$$
\varepsilon_{yx} = \frac{1}{2} \gamma_{yx} = \frac{1}{2} \left( \frac{\partial u}{\partial z} + \frac{\partial w}{\partial x} \right) = \frac{1}{2} \left( \frac{\partial w}{\partial x} + \phi \right)
$$

(4)

### 2.2 Constitutive equations for the CET

The known constitutive equations for an elastic isotropic Timoshenko beam within considering Poisson effect are:

$$
\sigma_{xx} = E \varepsilon_{xx}
$$

(5)

$$
\tau_{yx} = G \gamma_{yx} = 2G \varepsilon_{yx}
$$

(6)

Where $E$ and $G$ are the Young and shear modulus respectively.

### 2.3 Additional kinematic properties for the MCST

For an elastic isotropic Timoshenko beam, the non-zero component of the rotation vector can be expressed as follows [2,15]:

$$
\theta_x = -\frac{1}{2} \left( \frac{\partial w}{\partial x} + \phi \right)
$$

(7)

Then, the non-zero components of the curvature tensor result in:

$$
\chi_{xx} = \chi_{yy} = \frac{1}{2} \left( \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial \theta_y}{\partial x} \right) = -\frac{1}{4} \left( \frac{\partial \phi}{\partial x} + \frac{\partial^2 w}{\partial x^2} \right)
$$

(8)

### 2.4 Additional constitutive equations for the MCST

The non-zero components of the couple stress tensor result:

$$
m_{yx} = 2Gl^2 \chi_{xy} = 2Gl^2 \chi_{yx} = m_{yx}
$$

(9)
where, \( l \) is the length scale parameter of the beam’s material.

3 GOVERNING EQUATIONS

The Hamilton’s principle can be used to establish the weak form of the governing equations for free vibration of a Timoshenko micro beam:

\[
\int_0^L \delta(T - U) \, dx
\]  

(10)

Basis on the MCST, the strain energy for a continuum body occupying a volume \( V \) can be expressed as follows [1,15,17]:

\[
U = \frac{1}{2} \int_V \left( \sigma_{ij} \varepsilon_{ij} + m_y \chi_y \right) \, dV
\]  

(11)

where, the first term in the Eq. (11) constitutes the classical strain energy density and the second term constitutes the additional strain energy density, produced by the couple stress components. The components \( m_y \) form the symmetric part of the couple stress tensor and the components \( \chi_y \) form the symmetric part of the curvature tensor.

In the particular case of a Timoshenko micro beam model, the mentioned expression of the strain energy is expressed as follows:

\[
U = \frac{1}{2} \int_0^L \left( \sigma_{xx} \varepsilon_{xx} + 2\tau_{xz} \varepsilon_{xz} + 2m_y \chi_{xx} \right) \, dA \, dx
\]  

(12)

After replacing Eqs. (3-9) in Eq. (12) and integrating along the cross section of the beam, it is possible to express the strain energy in terms of the displacements in the following form:

\[
U = \frac{1}{2} \int_0^L \left[ EI \left( \frac{\partial \phi}{\partial x} \right)^2 + \kappa GA \left( \frac{\partial \omega}{\partial x} - \phi \right)^2 + \frac{I^2 GA}{4} \left( \frac{\partial^3 \omega}{\partial x^3} + \frac{\partial^3 \phi}{\partial x^3} \right) \right] \, dx
\]  

(13)

where \( \kappa \) is the shear factor, \( I \) and \( A \), are the moment of inertia and the cross sectional area.

The kinetic energy is not modified by the MCST, so it remains equal as in the CET.

In terms of time-dependent displacements it is given by:

\[
T = \frac{1}{2} \int_0^L \left( \rho A \left( \frac{\partial \omega}{\partial x} \right)^2 + \rho I \left( \frac{\partial \phi}{\partial x} \right)^2 \right) \, dx
\]  

(14)

where \( \rho \) is the mass density of the material of the beam. These functions can be expressed as:

\[
\omega(x,t) = w(x) e^{i\omega t}
\]

(15)

\[
\phi(x,t) = \phi(x) e^{i\omega t}
\]

(16)

After eliminating the time variable \( t \), it results:

\[
T = \frac{\omega^2}{2} \int_0^L \left( \rho A w^2 + \rho I \phi^2 \right) \, dx
\]  

(17)

where, \( \omega \) are the natural frequencies of the system. From Eq. (10) it follows:
Replacing Eqs. (13) and (17) in Eq. (18) the weak form of the governing equations is obtained:

\[
\delta(T - U) = 0
\]

(18)

\[
\int_0^L \left( EI \frac{\partial^2 \phi}{\partial x^2} + \kappa GA \left( \frac{\partial w}{\partial x} - \phi \right) \left( \frac{\partial \delta w}{\partial x} - \frac{\partial \phi}{\partial x} \right) \right) dx + \int_0^L \left( \frac{E I}{4} \frac{\partial^2 \phi}{\partial x^2} - \frac{E I}{2} \frac{\partial \phi}{\partial x} \left( \frac{\partial^2 \phi}{\partial x^2} + \frac{\partial \phi}{\partial x} \right) \right) dx - \omega^2 \int_0^L \left( \rho A w \delta w + \rho I \phi \delta \phi \right) dx = 0
\]

(19)

4 THE FINITE ELEMENT METHOD

The finite element method (FEM) has been used to solve Eq. (19), by implementing the Timoshenko beam element based on the MCST, developed by Kahrobaiyan et al [17]. This is a two node element that consists in a generalization of the classic cubic-quadratic Timoshenko beam element, which takes into account the size dependent material properties. The Figure 2 shows a scheme of the mentioned element.

The corresponding nodal displacements vector is:

\[
\{u_n\} = \{w_1 \phi_1 w_2 \phi_2\}^T
\]

(20)

Defining the non-dimensional space variable as follows:

\[
\xi = x / a
\]

(21)

Where, \(a\) is the length of the element, the followings are the corresponding shape functions for \(w\) and \(\phi\):
Where, \( \varphi \) is the non dimensional coefficient of the material properties and it is expressed as:

\[
\varphi = \frac{12EI}{\kappa G A a^2} (1 + \alpha)
\]  

(23)

and, \( \alpha \) is the non dimensional coefficient of the length scale parameter.

\[
\alpha = \frac{G A l^2}{EI}
\]

(24)

The generalized displacement vector of the element takes the form:

\[
\begin{bmatrix}
[w] \\
[\phi]
\end{bmatrix} =
\begin{bmatrix}
[N^e] \\
[N^s]
\end{bmatrix}
\begin{bmatrix}
[w_1] \\
[w_2] \\
[\phi_1]
\end{bmatrix}
\]

(25)

Directly, from the Eqs. (13) and (17) the strain energy and the kinetic energy of each element can be obtained as:

\[
U_e = \frac{1}{2} \{u_e\}^T [K_e] \{u_e\} = \frac{1}{2} \int_0^l \left( EI \left( \frac{\partial \phi}{\partial x} \right)^2 + \kappa G A \left( \frac{\partial w}{\partial x} - \phi \right)^2 + \frac{l^2 G A}{4} \left( \frac{\partial^2 w}{\partial x^2} + \frac{\partial \phi}{\partial x} \right)^2 \right) dx
\]

(26)

\[
T_e = \frac{1}{2} \{u_e\}^T [M_e] \{u_e\} = \frac{\alpha^2}{2} \int_0^l \left( \rho A w^2 + \rho I \phi^2 \right) dx
\]

(27)

where, \([K_e] \) and \([M_e] \) are the stiffness and mass matrices of each element, respectively.

Replacing the Eq. (25) in the Eqs. (26-27), the expressions of the stiffness and mass matrices can be obtained as follows:

\[
[K_e] = \int_0^l \left( EI \left( \frac{N^{e\varphi}}{a} \right)^T \frac{N^{e\varphi}}{a} + \kappa G A \left( \frac{N^{e\varphi}}{a} - N^{e\phi} \right)^T \left( \frac{N^{e\varphi}}{a} - N^{e\phi} \right) \right) a d \xi
\]

(28)

\[
[M_e] = \omega^2 \int_0^l \left( \rho A \left( N^{e\varphi} \right)^T N^{e\varphi} + \rho I \left( N^{e\phi} \right)^T N^{e\phi} \right) a d \xi
\]

(29)

Eqs. (28-29) have been implemented in a finite element algorithm and by means of a convergence analysis, a mesh with 50 elements was adopted to obtain the numerical results.
4 NUMERICAL RESULTS

Three combinations of classical boundary conditions of micro beams are analyzed using the proposed finite element algorithm. Figure 3 shows the three cases under study.

![Figure 3: Timoshenko micro-beam models under study, with different classical boundary conditions: a) Pinned-pinned, b) Clamped-free, c) Clamped-clamped.](image)

In all the cases, the ratio between the natural frequencies $\omega_{n}^{\text{MCST}}$, obtained by the MCST and the natural frequencies $\omega_{n}^{\text{CET}}$, obtained by the CET, has been plotted. The non dimensional length scale ratio $h/l$, has been ranged from 0.2 to 20 and to better appreciated the results, the corresponding ranges of $h/l$ have been divided in two different graphics, so in each one of them, figure “a” corresponds to $h/l \leq 1$ and the figure “b” corresponds to $h/l \geq 1$. A rectangular cross section with an aspect ratio $L/h = 10$ has been adopted. Since these graphics have been plotted with non dimensional parameters, they are independent of the material properties.

Additional graphics have been plotted with dimensional parameters, for the particular case of epoxy material with the following properties:
- Material length scale parameter $l = 17.6 \mu m$.
- Young modulus $E = 1.44 GPa$.
- Mass density $\rho = 1250 Kg/m^3$.

Each figure contains from 1 up to 3 different curves. The black curves correspond to exact results with MCST, obtained from expressions available in the literature [9], this type of curves appears only for the pinned-pinned boundary condition. The red curves correspond to MCST results, obtained with FEM and the blue dashed curves correspond to CET results.

The first case under study corresponds to a pinned-pinned Timoshenko micro-beam.

![Figure 4: First natural frequency ratio $\omega_{1}^{\text{MCST}} / \omega_{1}^{\text{CET}}$ for a pinned-pinned Timoshenko micro-beam: a) In the range $h/l \leq 1$, b) In the range $h/l \geq 1$.](image)

In this case, the values obtained with FEM are in good agreement with the exact solution.
when $h/l \geq 1$. By other hand, the CET is not accurate when $h/l \leq 20$. In Figure 4, the first natural frequency ratio $\omega_{1}^{\text{mcst}} / \omega_{1}^{\text{cet}}$ for a pinned-pinned Timoshenko micro-beam has been plotted and in Figure 5 the corresponding frequencies $f_{1}^{\text{mcst}}$ and $f_{1}^{\text{cet}}$ have been plotted, for the case of an epoxy material.

![Figure 5](image1)

**Figure 5**: First natural frequency $f_{1}^{\text{mcst}}$ and $f_{1}^{\text{cet}}$, for a pinned-pinned Timoshenko micro-beam, for epoxy material: a) In the range $h/l \leq 1$, b) In the range $h/l \geq 1$.

It is very important to mention that, the experimental values (not obtained in the present work) are in good agreement with those obtained with the MCST, in all of the cases found in the available literature. In Figure 6 and 7 has been plotted the second natural frequency for a pinned-pinned Timoshenko micro-beam.

![Figure 6](image2)

**Figure 6**: Second natural frequency ratio $\omega_{2}^{\text{mcst}} / \omega_{2}^{\text{cet}}$, for a pinned-pinned Timoshenko micro-beam: a) In the range $h/l \leq 1$, b) In the range $h/l \geq 1$.

![Figure 7](image3)

**Figure 7**: Second natural frequency $f_{2}^{\text{mcst}}$ and $f_{2}^{\text{cet}}$, for a pinned-pinned Timoshenko micro-beam, for epoxy material: a) In the range $h/l \leq 1$, b) In the range $h/l \geq 1$. 

a) MCST exact
b) MCST with FEM

$\omega_{2}^{\text{mcst}} / \omega_{2}^{\text{cet}}$

$\omega_{2}^{\text{mcst}}$

$\omega_{2}^{\text{cet}}$

(a) h/l

(b) $h/l$

$\omega_{2}^{\text{mcst}}$

$\omega_{2}^{\text{cet}}$

(a) $h/l$

(b) $h/l$

$\omega_{2}^{\text{mcst}}$

$\omega_{2}^{\text{cet}}$

(a) $h/l$

(b) $h/l$
It can be seen that for the second natural frequency the results obtained with FEM are in good agreement too, with the exact solution, when $h/l \geq 1$. The second case corresponds to a clamped-free Timoshenko micro-beam. In Figure 8, the first natural frequency ratio $\omega_1^{mcst} / \omega_1^{cet}$ for a clamped-free Timoshenko micro-beam has been plotted and in Figure 9 the corresponding frequencies $f_1^{mcst}$ and $f_1^{cet}$, are plotted for the epoxy material in consideration. In the graphics corresponding to epoxy material, the frequencies are expressed in $MHz$.

Figure 8: First natural frequency ratio $\omega_1^{mcst} / \omega_1^{cet}$, for a clamped-free Timoshenko micro-beam: a) In the range $h/l \leq 1$, b) In the range $h/l \geq 1$.

Figure 9: First natural frequency $f_1^{mcst}$ and $f_1^{cet}$, for a clamped-free Timoshenko micro-beam, for epoxy material: a) In the range $h/l \leq 1$, b) In the range $h/l \geq 1$.

In Figure 10, the second natural frequency ratio $\omega_2^{mcst} / \omega_2^{cet}$ for a clamped-free Timoshenko micro-beam has been plotted and in Figure 11 the corresponding frequencies $f_2^{mcst}$ and $f_2^{cet}$, are shown too, for the epoxy material in consideration.

Figure 10: Second natural frequency ratio $\omega_2^{mcst} / \omega_2^{cet}$, for a clamped-free Timoshenko micro-beam: a) In the range $h/l \leq 1$, b) In the range $h/l \geq 1$. 

454
As expected, the natural frequencies $f_{n}^{mcst}$, tends to $f_{n}^{cet}$ when $h \gg l$. The last case under study corresponds to a clamped-clamped Timoshenko micro-beam.

In Figure 12, the first natural frequency ratio $\omega_{1}^{mcst} / \omega_{1}^{cet}$ for a clamped-clamped Timoshenko micro-beam has been plotted and in Figure 13 the corresponding frequencies $f_{1}^{mcst}$ and $f_{1}^{cet}$, are shown again, for the particular case of an epoxy material.

In Figure 14, the second natural frequency ratio $\omega_{2}^{mcst} / \omega_{2}^{cet}$ for a clamped-clamped condition has been plotted and in Figure 15 the corresponding frequencies $f_{2}^{mcst}$ and $f_{2}^{cet}$, are
shown too, for the epoxy material.

**Figure 14:** Second natural frequency ratio $\frac{\omega_{2}^{\text{mcst}}}{\omega_{2}^{\text{cet}}}$, for a clamped-clamped Timoshenko micro-beam: a) In the range $h/l \leq 1$, b) In the range $h/l \geq 1$.

**Figure 15:** Second natural frequency $f_{2}^{\text{mcst}}$ and $f_{2}^{\text{cet}}$, for a clamped-clamped Timoshenko micro-beam, for epoxy material: a) In the range $h/l \leq 1$, b) In the range $h/l \geq 1$.

## 5 CONCLUSIONS

- In the present numerical analysis, it is clear that the results obtained by the CET differ significantly from those results calculated by the MCST, especially in the range $h/l \leq 20$.

- In the particular case of pinned-pinned boundary condition, the results obtained with the MCST, using FEM, are in good agreement with the exact solution obtained with the MCST by Araújo et al [9], in the range $h/l \geq 1$.

- It is important to mention that from experimental results in the available literature, the present results based on MCST are in very good agreement with them.

## ACKNOWLEDGEMENTS

The present study has been sponsored by the Universidad Nacional del Sur, UNS, the Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET and the Comisión de Investigaciones Científicas, Buenos Aires Province (CIC). It was developed in the Departamento de Ingeniería and Instituto de Mecánica Aplicada, IMA, UNS.
REFERENCES


A STRESS ANALYSIS APP

SERGIO GALLEGOS-CAZARES, ALEJANDRA BALDERRAMA-LOPEZ

Tecnologico de Monterrey, ENCIT, Campus Monterrey,
Av. Eugenio Garza Sada 2501, Engineering Laboratories Building, 206K, Monterrey N.L. Mexico
64849
sergio.gallegos@itesm.mx

Key words: tablet computing, engineering education, engineering apps.

Summary. Understanding abstract engineering concepts is difficult for engineering students because most of them are of the visual type. Current mobile technology combined with computational mechanics offers innovative ways to create computational tools that may ease the comprehension of these concepts. This work explores some ideas behind the development of an iOS app to simulate a stress analysis laboratory to be used in an educational environment. Using iPad devices a library of engineering mechanics experiments can be manipulated. Engineering students can visualize stresses and displacements using gestures to manipulate stress and displacement contours, scale colors, and query values at a point. Modification of an experiment triggers a reanalysis using finite elements. Several families of experiments are contained in this app simulating bars, beams, frames and stress concentration problems. Current mobile devices are adequate platforms for application of computational mechanics and offer an opportunity for the development of a new set of educational tools in engineering.

1 INTRODUCTION

Several authors have studied and modeled the learning styles of engineering students. Accordingly with the Felder-Silverman model [1, 2] there are four dimensions characterizing different learning styles: Sensory-intuitive, visual-verbal, active-reflective and sequential-global. The first refers to the type of information that the student preferentially perceives; the sensory student prefers sights, sounds, and physical sensations while the intuitive prefer memories, thoughts and insights. The second is related to the type of information that is most effectively perceived; visual students perceive pictures, diagrams, flow charts and demonstrations while the verbal are good with written and spoken explanations. The third defines how the students prefer to process information; active students prefer engagement in physical activity or discussion while the reflective prefer introspection. The fourth dimension characterizes how the student progress toward understanding; the sequential prefers a logical progression of incremental steps while the global does it in big picture jumps. From a study [2] of 2506 engineering students in 2005 it was found that approximately 63% are sensory, 80% are visual, 64% are active and 60% are sequential.

Traditional teaching styles in engineering have been mostly theoretical, verbal and deductive with students watching and listening. There is a mismatch between the learning style of most students and the traditional teaching style. A more active teaching role has been
promoted in the universities in order to improve student satisfaction and reduce the number of dropouts. Mobile learning [3] uses Internet, sensors and devices to allow any time, any place interaction and it has been shown [4] to promote the development the cognitive skills in students. Tablet computing [5] is a key technological trend that is gaining momentum in education because the students can download the apps of their interest making their devices portable personalized learning environments. Current tablets use large screens, gesture based interfaces, better processors, connectivity and access to many types of sensors creating a new technology in its own right. The development of adequate mobile apps tailored to engineering student needs is essential for the success of tablet computing as an educational tool.

The objective of this research is to present the characteristics of a tablet application that promotes active learning activities and improves the understanding of solid mechanics concepts for engineering students. The application allows the students to select among a set of mechanics experiments, visualize deformation, different types of stresses, query the values using gestures and modify the type of loading to see the change on the experiment.

The application is developed to work in an iPad using the iOS operating system and the Objective-C language. An object oriented finite element model is created to allow description of experiments, manipulation of mechanical data and in order to make reanalysis possible after a load change.

The application works in iPhones and iPads, but the focus here is in the iPad platform due to the larger screen real-state that allows for better graphics and simplifies the querying or selection of data. The graphics are drawn using OpenGL and due to the high resolution of the Retina display are of great quality. A number of experiment models are included modeling bars, beams, frames and stress concentration problems. If the loading pattern is changed a reanalysis of the model is possible due to the finite element model behind the experiment, but the analysis times are strongly dependent on the type of processor in the device.

The development of the iPad stress laboratory describes how to combine tablet technology and computational mechanics to produce an educational tool that may improve the understanding of abstract engineering concepts by allowing student interaction in a sensory, active and visual style. Finite element technology allows of data handling and reanalysis of computational models in these devices but computational times are dependent on the type of processor and may take more time than that required for interactive use of the application.

2 IPAD IMPLEMENTATION

2.1 Finite elements with Objective-C

All the experiments in the computational laboratory are modeled using plane stress triangular finite elements. Since the target implementation platform was an iPad, a finite element program named Xplore was developed using Objective-C, which is the language of implementation for iOS [6, 7]. Objective-C is an object oriented, superset of C and dynamically typed language that allows great flexibility in the representation of complex objects. Xplore models a finite element model using a mesh object to capture the geometry, connectivity and properties of the elements, and a step object to model the boundary conditions and the solution. When an Xplore solver method is called, it takes the mesh object and the boundary conditions to compute the step solution for the current step. From this point
of view, an analysis is just a set of step objects related to a mesh. A complete finite element environment was assembled as an independent MAC application using GID [8] for pre and post processing with Xplore as an analysis tool. Each experiment was modeled and solved separately. For each experiment the mesh and the step objects were stored in a file using a property list format [9] or plist format.

2.2 App structure

The flow of information is described in figure 1. User interaction starts by selecting an experiment on the screen from a list. A Table View controller [10], named SL2ExperimentViewController, receives the action and passes the request to the application store. The store chooses the right finite element model from a catalog of experiments and returns it back to the controller. Finally, the controller will be responsible to transform the experiment into a graphics model to be displayed on the screen. The store is the object handling all requests for models and is a unique instance called a singleton [11].

![Figure 1: Flow of information for the application.](image)

The app interface uses a split view controller as shown figure 2 with a catalog of experiments to the left and a graphics screen to the right. The experiments are divided in groups for axially loaded bars, beams, frames and stress concentration problems. For each experiment a general description, the geometry, the boundary conditions and the material employed are presented upon request as shown in figure 3.

Each experiment is a complete finite element model, which is loaded as a resource into the app using a plist format. The store uses a directory of the available experiments, which is also loaded as a resource, to locate an individual experiment.
Figure 2: User interface with a catalog of experiments to the left and a graphics screen to the right.

Figure 3: Detailed information about a selected experiment.
2.3 Graphics model

Once the main controller receives the finite element model, it must create a graphics object of the class GLModel to display it on the screen. This class creates a mesh of triangles to describe the geometry and stores the required information at the nodes. The controller will then use OpenGL commands to display the graphics. There are three graphics states: test, stress and displacement that can be selected with a switch. The default state is test where the controller displays a texture, the loads with blue markers and the displacement constraints with red markers, as shown in figure 2. When the switch is used to select stress or displacement, the controller displays contour plots. For example, in figure 4 the von Mises stress contours are displayed for one experiment. Manipulation of the graphics information is achieved by using gestures. The figure can be reduced or amplified by pinching on the screen; translation is achieved by pressing with one finger and moving it; and querying the value at a point is achieved by simply pointing with the finger at the desired position using a long pressure gesture. In addition, tapping on the scale changes the number of colors used to plot from 300 to 2 producing some interesting effects. Different types of contours may be selected: von Mises, horizontal and vertical stresses, shear stress, maximum shear, horizontal and vertical displacements or magnitude of displacement. Tapping on top of the model changes the type of contour drawn.

![Figure 4: Von Mises stresses in a plate with a circular hole.](image-url)
2.4 Reanalysis

Xplore is an object oriented finite element code that models the mesh as an object separate from the step object, which contains the boundary conditions and results. If the boundary conditions change, Xplore’s solver method is called to generate a new set of step results. From this point of view, an analysis is just a set of step objects representing the change in the state of the mesh object.

When an experiment is in the test mode, the boundary conditions are presented with blue markers for loads and red markers for displacement constraints, as shown in figure 2. Applying a long pressure gesture near a load marker a pop up form is presented to the user, allowing edition. Once a change in boundary conditions is detected, an internal flag is turned on. When the stress or displacement mode is selected and the flag is on, a new step object is created and the solver method is called to update the step solutions, reflecting the current state of the model.

3 RESULTS

3.1 Axially loaded bars

Five bar experiments where introduced in the application. Figures 5 (a) to (d) show von Mises stresses in a single bar with different aspect ratios ranging from thick to slender. Figure 5 (e) presents two bars with a change in transverse sectional area.

Figure 5: Von Mises stresses in axially loaded bars. (a) L/h=2. (b) L/h=3. (c) L/h=7.5. (d) L/h=20. (e) Two bars.
3.2 Beams

Two beam experiments are included in the application. Figure 6 shows the normal longitudinal stress for: (a) a fixed-fixed beam under uniform vertical load, and (b) a simply supported beam with concentrated vertical loads at the thirds.

![Figure 6: Longitudinal stress. (a) Fixed-fixed beam with uniform load. (b) Simply supported beam with two concentrated loads.](image)

3.3 Frames

Three frame experiments are included in the laboratory as shown in figure 7. Von Mises stresses are presented for (a) a fixed frame under uniform vertical load, (b) a fixed frame under lateral load, and (c) an L-frame fixed at the base and with a concentrated load at the tip.


### Figure 7: Von Mises stress in frames.

(a) Fixed-fixed frame with uniform load. (b) Fixed-fixed frame with lateral load. (c) L Frame with a concentrated load at the tip.

### 3.4 Stress concentration

Two stress concentration experiments are included in the laboratory. Figure 8 presents two plates under a horizontal tensile field with (a) a circular hole, and (b) an elliptical hole in the middle.

### Figure 8: Von Mises stress in plates under tension.

(a) Circular hole. (b) Elliptical hole.
3.5 Time for reanalysis

After a change in the boundary conditions a reanalysis of the finite element model is triggered for the current analysis step. Times where measured in an iPad 4 from the moment the stress state was selected with the state switch until the von Mises contours are drawn on the screen. Results are reported in table 1 and compared to the number of degrees of freedom (DOFs) and the number of elements.

Table 1: Times measured in an iPad 4 for reanalysis and contour display of each model.

<table>
<thead>
<tr>
<th>Model</th>
<th>Reanalysis time</th>
<th>DOFs</th>
<th>Elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bar with L/h = 7.5</td>
<td>20 seg</td>
<td>1262</td>
<td>1106</td>
</tr>
<tr>
<td>Bar with L/h =3</td>
<td>1 min 06 seg</td>
<td>1990</td>
<td>1877</td>
</tr>
<tr>
<td>Bar with L/h =2</td>
<td>1 min 19 seg</td>
<td>2128</td>
<td>2038</td>
</tr>
<tr>
<td>Two Bars</td>
<td>17 seg</td>
<td>956</td>
<td>831</td>
</tr>
<tr>
<td>Bar with L/h =20</td>
<td>36 seg</td>
<td>2158</td>
<td>1844</td>
</tr>
<tr>
<td>Fixed beam</td>
<td>7 seg</td>
<td>532</td>
<td>442</td>
</tr>
<tr>
<td>Beam with 2 loads</td>
<td>48 seg</td>
<td>2212</td>
<td>1974</td>
</tr>
<tr>
<td>Frame with horizontal load</td>
<td>1 min 27 seg</td>
<td>2576</td>
<td>2268</td>
</tr>
<tr>
<td>L frame</td>
<td>1 min 47 seg</td>
<td>3242</td>
<td>2952</td>
</tr>
<tr>
<td>Frame with vertical load</td>
<td>59 seg</td>
<td>2070</td>
<td>1713</td>
</tr>
<tr>
<td>Plate with a circular hole in tension</td>
<td>1 min 23 seg</td>
<td>1472</td>
<td>1328</td>
</tr>
</tbody>
</table>

4 CONCLUSIONS

A stress laboratory is a personalized application that allows visualization of abstract concepts like stress that cannot be visualized in any other way. In addition, these applications allow visualization of concrete concepts like displacement, but enhance their comprehension by amplifying it. Tablet devices, and particularly the iPad, are adequate tools for the development of these applications and may enhance self-learning and teaching processes.

Sets of simple experiments can be packed inside such an application and organized by topics allowing active teaching strategies to be used in class. Particularly, such an application would be very useful if the experiments illustrate common topics studied in basic solid mechanics courses.

Tablets are very appealing devices for this type of applications. Use of gestures to manipulate the graphics display and the types of results presented make apps fast and responsive. The high resolution of the screen generates high quality and attractive graphics contours that illustrate mechanical behavior very clearly for visual learners.

Tablets can be used for finite element computations. Reanalysis is performed automatically when changes in the model are generated; however, total times for analysis and contour drawing are not small enough for the application to have an immediate response to the user interaction. Exploration of alternative finite element models organization and of alternatives to transform finite element information into graphics models need to be done to improve speed of response of apps in tablets.

Future work must address alternative graphics display using vectors, improve the graphics representation of boundary conditions, improve the experiment library with other families of
problems and more experiments for each family. Further research on better organization of
finite element models and transformation to graphics models for tablets is needed to improve
response time. Use of newer tablet platforms, like the iPad Air, with better and faster
processors may also improve response times. Finally, extension to iPhones or smart
telephones may also help to make this type of applications available to a larger number of
students, even at the expense of a smaller real-state screen.

REFERENCES
    New Media Consortium.
    Consortium.
    (2015)
    Wesley, (2013).
[8] CIMNE. *GID, the personal pre and post processor*. International Center for Numerical
[9] iOS Developer Library. *Property list programming guide*, Apple online documentation
[10] iOS Developer Library. *Table View programming guide for iOS*, Apple online
    4th edition (2014), Big Nerd Ranch, LLC.
USE OF COMPUTATIONAL FLUID DYNAMIC TO COMPARE THE PRESSURE LOSS BETWEEN A PARALLEL FLOW FIELD PLATE AND A PARALLEL-BAFFLE FLOW FIELD PLATE IN A DIRECT ETHANOL PROTON EXCHANGE MEMBRANE FUEL CELL

J. VICENTE¹, G. BENETTI¹, M. FORTE², G. HEUSY¹, E. MATHIAS¹,

P. BELCHOR¹, ²

¹UNOESC (Brazil)
campus II of Joaçaba – Rua José Firmo Bernardi, Joaçaba (Brazil)
jonatas.vicente7@yahoo.com.br / http://www.unoesc.edu.br/

²UFRGS (Brazil)
campus do Vale – Av. Bento Gonçalves, 9500, Set. 4, Pr. 74, Porto Alegre (Brazil)
rafrstv@hotmail.com / http://www.ufrgs.br/

Key words: Fuel Cell, CFD, Flow Field Plate, Pressure Loss.

Abstract. This research provided CFD tests to verify and understand the reagents fluid dynamic behavior in the channels of the new parallel-baffle flow field plate during a direct ethanol proton exchange membrane fuel cell (DE-PEMFC) operation. The results showed that the new flow field plate is able to concentrated all pressure loss in an unique pathway, improving the capacity of expel water produced in the cathode and decreasing the pressure loss effects in both DE-PEMFC electrodes, improving the fuel cell overall performance.

1 INTRODUCTION

Fuel cells are considered to be the green power sources for the 21st century, and may make the “hydrogen economy” a reality [1]. The fuel cell was discovered around two centuries ago by William Grove and Friedrich Schönbein and has yet to see widespread commercialization, despite the advantages of the technology and its wide range of potential applications. The reason for the delay in commercialization of the technology can be ascribed to several present characteristics of the fuel cell: for example, cost and complexity, immaturity, and its role as replacement technology [2]. Improvements in cell design and manufacturing have further increased power, while reducing manufacturing costs, which is essential if the fuel cell is to compete with the internal combustion engine [3].

According to Li et al. [3], bipolar plates comprise about 60% of the weight and 30% of the total cost in a fuel cell stack. The channels design and its pattern considerably affect the effectiveness of mass transport as well as electrochemical reactions inside the cell. More recently, there was a trend to apply CFD methods to fuel cells modeling [4].

Computational fluid dynamics (CFD) modeling is the most common approach to creating simulations of ethanol alcohol within a direct ethanol proton exchange membrane fuel cell
(DE-PEMFC) [5]. The pressure loss in the DE-PEMFC flow field plates decrease the power density and the fuel cell performance. Within parallel flow channels the reactant gas speed is low relative to serpentine flow fields which may lead to local flooding, particularly under the lands, where the pressure gradient is a minimum [6].

The basic idea of interdigitated flow fields is to force the total mass flow through the land area to improve the local cell performance [7]. However, here the excessive pressure drop from the gas inlet to the gas outlet of the flow field requires additional parasitic power [8].

2 OBJECTIVES

The aim of this research was project and characterized the pressure loss behavior related to a new DE-PEMFC flow field design, with parallel flow field plates (PFFP) and interdigitated flow field plates (IFFP) characteristics, a parallel-baffle flow field plate (PBFFP).

3 METHODOLOGY

The new flow field plate design were created by SOLIDWORKS software 2013. The CFD tests was employed by SOLIDWORKS flow simulation tool in a computer model Alienware Aurora Desktop – BRH3171 (3.2 GHz, 8 MB L3 cache; 24GB DDR3 1333MHz memory (6x4GB)) with an high-performance liquid cooling (Alienware®), equipped with a Intel® Core™ i7-960. As comparison, in the simulations both flow plates, the classic PFFP and the new PBFFP, showed in the Figure 1, received a volume flow of 1 L/min (hydrogen) in the inlet and an environment pressure in the outlet.

Figure 1: (a) the new PBFFP and (b) the PFFP
4 RESULTS AND DISCUSSIONS

The result showed that in the PFFP all channels suffered with pressure loss, as showed in the Figure 2.

Figure 2: Pressure loss behavior in the PFFP channels.

In the PBFFP, the interdigitated channels pressure loss was stabilized and all pressure loss was concentrated in the unique connecting pathway between the inlet and the outlet, as shown in the Figure 3.

Figure 3: Pressure loss behavior in the new PBFFP channels.

The Table 1 summarizes all the variables related to Hydrogen fluid dynamic behavior in the PFFP.
Table 1: Variables related to Hydrogen fluid dynamic behavior in the PFFP

<table>
<thead>
<tr>
<th>Name</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pressure [Pa]</td>
<td>101324.93</td>
<td>101347.59</td>
</tr>
<tr>
<td>Temperature [K]</td>
<td>298.15</td>
<td>373.15</td>
</tr>
<tr>
<td>Density [kg/m^3]</td>
<td>0.07</td>
<td>0.08</td>
</tr>
<tr>
<td>Velocity [m/s]</td>
<td>0</td>
<td>10.739</td>
</tr>
<tr>
<td>Velocity (X) [m/s]</td>
<td>-10.739</td>
<td>2.241</td>
</tr>
<tr>
<td>Velocity (Y) [m/s]</td>
<td>-9.026</td>
<td>1.740</td>
</tr>
<tr>
<td>Velocity (Z) [m/s]</td>
<td>-2.685</td>
<td>2.711</td>
</tr>
<tr>
<td>Temperature (Fluid) [K]</td>
<td>298.15</td>
<td>373.15</td>
</tr>
<tr>
<td>Temperature (Solid) [K]</td>
<td>372.80</td>
<td>373.15</td>
</tr>
<tr>
<td>Mach Number [ ]</td>
<td>0</td>
<td>7.32e-003</td>
</tr>
<tr>
<td>Vorticity [1/s]</td>
<td>2.198</td>
<td>51038.952</td>
</tr>
<tr>
<td>Shear Stress [Pa]</td>
<td>1.83e-005</td>
<td>1.62</td>
</tr>
<tr>
<td>Relative Pressure [Pa]</td>
<td>-0.07</td>
<td>22.59</td>
</tr>
<tr>
<td>Heat Transfer Coefficient [W/m^2/K]</td>
<td>7.405e-013</td>
<td>2367.448</td>
</tr>
<tr>
<td>Surface Heat Flux [W/m^2]</td>
<td>-0.234</td>
<td>188678.372</td>
</tr>
<tr>
<td>Heat Flux [W/m^2]</td>
<td>0</td>
<td>146974.586</td>
</tr>
<tr>
<td>Overheat above Melting Temperature [K]</td>
<td>-3627.348</td>
<td>-3627.000</td>
</tr>
</tbody>
</table>

The Table 2 summarizes all the variables related to Hydrogen fluid dynamic behavior in the PBFFP.
Table 2: Variables related to Hydrogen fluid dynamic behavior in the PBFFP

<table>
<thead>
<tr>
<th>Name</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pressure [Pa]</td>
<td>101324.93</td>
<td>101398.45</td>
</tr>
<tr>
<td>Temperature [K]</td>
<td>298.15</td>
<td>373.15</td>
</tr>
<tr>
<td>Density [kg/m^3]</td>
<td>0.07</td>
<td>0.08</td>
</tr>
<tr>
<td>Velocity [m/s]</td>
<td>0</td>
<td>11.165</td>
</tr>
<tr>
<td>Velocity (X) [m/s]</td>
<td>-10.928</td>
<td>10.924</td>
</tr>
<tr>
<td>Velocity (Y) [m/s]</td>
<td>-10.939</td>
<td>2.702</td>
</tr>
<tr>
<td>Velocity (Z) [m/s]</td>
<td>-2.955</td>
<td>2.943</td>
</tr>
<tr>
<td>Temperature (Fluid) [K]</td>
<td>298.15</td>
<td>373.15</td>
</tr>
<tr>
<td>Temperature (Solid) [K]</td>
<td>372.80</td>
<td>373.15</td>
</tr>
<tr>
<td>Mach Number [ ]</td>
<td>0</td>
<td>7.61e-003</td>
</tr>
<tr>
<td>Vorticity [1/s]</td>
<td>1.971e-007</td>
<td>72074.086</td>
</tr>
<tr>
<td>Shear Stress [Pa]</td>
<td>0</td>
<td>1.68</td>
</tr>
<tr>
<td>Relative Pressure [Pa]</td>
<td>-0.07</td>
<td>73.45</td>
</tr>
<tr>
<td>Heat Transfer Coefficient [W/m^2/K]</td>
<td>0</td>
<td>2367.207</td>
</tr>
<tr>
<td>Surface Heat Flux [W/m^2]</td>
<td>-1.206</td>
<td>188657.930</td>
</tr>
<tr>
<td>Heat Flux [W/m^2]</td>
<td>0</td>
<td>147161.252</td>
</tr>
<tr>
<td>Overheat above Melting</td>
<td>-3627.349</td>
<td>-3627.000</td>
</tr>
<tr>
<td>Temperature [K]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Pressure loss is one of the important parameters that affect the total system efficiency and optimization. On the other hand, a low pressure drop might inhibit liquid water removal, which can cause starvation of fuel or oxidant, potentially damaging the electrode and membrane.

Temperature, flow rates and pressure drop in a channel determine water accumulation both in the gas diffusion layer (GDL) and the channel [9-10].

Thus, the total pressure drop along the channel will increase the amount of carried water vapor and taken away by a given amount of the reactant gas flow if the relative humidity and temperature are maintained. Thus the relationship between water accumulation and remove is dependent on the flow rate and the pressure drop in a channel [11].

5 CONCLUSION

The result showed that in the PFFP all channels suffered with pressure loss but in the PBFFP, the interdigitated channels pressure loss was stabilized and it was concentrated only in the pathway connecting inlet and outlet. A flow field plate able to concentrate all pressure loss in a unique pathway could be used in a DE-PEMFC cathode side, keeping the homogeneous pressure and improving the water expelling, improving the overall DE-PEMFC performance.
6 ACKNOWLEDGEMENT

The financial support of Agência Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES/PROCAD Project nº 200/2007), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Future (Science and Technology) is gratefully acknowledged.

REFERENCES

HYDRAULIC FRACTURING SIMULATIONS USING A HOMOGENIZED CONTINUUM APPROACH

PEDRO H. E. CHEROTO

Engineering Simulation and Scientific Software
Avenida Presidente Vargas 3131, Cidade Nova, Rio de Janeiro, RJ, Brasil
Pedro.cheroto@esss.com.br

Key Words: Hydraulic Fracturing, Ansys, Homogenized Continuum Approach, Smeared Continuum Approach, Oil & Gas.

1 INTRODUCTION

The study of Hydraulic Fracturing and its numerical modelling present many difficulties. Onsite experimental tests for research and development, although extremely valuable (and necessary), are very expensive. Numerical modelling represents the possibility of doing this research for a fraction of the costs, and yet, such approach has its own problems. Perhaps the most limiting of those is the necessity to properly represent fracture propagation in a solid continuum; discrete modelling of this effect implies the use of very fine meshes which, on a three dimensional reservoir model, are unfeasible. Other issues include material modelling (correct characterization of soil and rock behavior, as well as anisotropic failure), difficulty of obtaining experimental data for input of material properties and boundary conditions, and the complexity of validation after the analysis is complete.

Progress in hydraulic fracturing, regardless of when in the chain of events, has continuously led to lower costs, and thus, higher income in the oil and gas industry. A clear example of this is the U.S.A., that thanks to hydraulic fracturing is expected to be the world’s top oil producer by 2020 [1]; not only is the U.S. reducing its costs by importing less, but also steadily generating the potential of being an international provider of oil.

That said, taking into consideration the difficulties aforementioned, most progress is being made by empirical and experimental means. Discrete modelling of fracture propagation, albeit the most obvious approach, has been a huge obstacle in the numerical modelling of this problem. An elegant alternative is the use of a homogenized continuum approach; in other words, a domain in which fractures are modelled at the material level. By doing so, one loses the ability to see individual fractures, but obtains reasonable global results – the actual desired output of a hydraulic fracturing simulation.

The present work uses a homogenized continuum approach to model two examples of hydraulic fracturing at a reservoir, with the objective of showing potential benefits and limitations of the method. Due to data confidentiality, material properties used in this paper will not be explicitly mentioned, and were altered from one case to another to provide different behavior during the fracking.
2 BRIEF INTRODUCTION TO THE METHOD

As mentioned before, the use of the homogenized continuum approach to simulate hydraulic fracturing is essential to reduce simulation time, but it models only the mechanical part of the analysis. Equally important are the approaches to model the fluid flow and the coupling of both physics, as well as the constitutive material models used to portray the material behavior. All of these topics will be briefly covered in this section and is explained in more detail by Will, Johannes [2].

2.1 Mechanical Analysis

In the smeared continuum approach (also known as Homogenized continuum approach), the deformation tensor of the jointed rock is composed from the actual rock stiffness and its multiple joint systems stiffness. As such, the stress strain condition of the jointed rock is given by:

$$\sigma = D_{RM}(\varepsilon^{tot} - \varepsilon^{nl})$$ (1)

Where $D_{RM}$ is the jointed rock deformation matrix resulted from the combined rock and joint stiffness $D_{RM} = D_{R} + D_{J}$.

In rock mechanics in general, the constitutive material models of elastic plastic behavior generally use effective stress:

$$\sigma_{tot} = \sigma_{eff} + np$$ (2)

Where

$$n = (1 \ 1 \ 1 \ 0 \ 0 \ 0)^T$$ (3)

And $p$ is the pore pressure.

2.2 Fracture Growth Constitutive Materials

To realistically model fracture growth in a reservoir, a jointed rock material model with capability of new joint generation is necessary. In other words, the material has a constitutive model for the “intact rock” and the constitutive model for the joints; the intersection of these material models introduces a concept known as Multisurface Plasticity. The creation of new joints is also necessary to take into account fracture propagation in any direction in the rock.

It must be emphasized that the mean joints are not modeled discretely, and that the strength conditions of the intact rock and joints are checked in every discretization point. That is, the homogenized continuum approach is used instead of a method using discrete fractures.

The software used to model this type of rock behavior was Ansys 14.5, in conjunction with Dynardo’s multiPlas, a library with which one can use such a material model in Ansys.

2.3 Fluid flow Analysis

Fluid flow occurs mainly in initiated or reopened joint systems due to the low permeability of the initial jointed rock. This is defined through an anisotropic permeability tensor of the homogenized rock mass. By using Darcy’s laminar flow approach, the joint system flow velocity in a determined joint set is:
Respectively transformed into global coordinate system \( \{V_{FT}\} = K[I] \)

\[
V_{FT} = \frac{q}{A} = k_{TF} \frac{2a_t}{d} I
\]

Where:
- \( V_{FT} \) = flow velocity
- \( K \) = anisotropic permeability tensor of jointed rock mass
- \( 2a_t \) = joint thickness / joint opening
- \( A \) = cross section
- \( d \) = joint frequency

3 PHYSICS COUPLING

Considering that the hydraulic fracturing simulation is actually a multiphysics event, it is necessary to provide a two-way coupling between the fluid flow and mechanical joint openings. That is done step by step during the analysis in the following way.

3.1 Fluid Flow to Mechanical Analysis

Immediately after the initial conditions for the analysis are generated (in situ stress and in situ pore pressure), the hydraulic fracturing simulation truly begins. For every fluid time increment, transient fluid flow affects the pore pressure field, which, in turn, generates an incremental change in the nodal forces of the mechanical analysis. These incremental changes of force are computed from the pore pressure gradients using the following relationship:

\[
\{F^{Str}\} = \gamma_W V_G \{I\}
\]

Where:
- \( F^{Str} \) = force vector
- \( I \) = gradients of pore pressure
- \( \gamma_W \) = water density
- \( V_G \) = related volume

3.2 Mechanical to Fluid Flow Analysis

At every mechanical step, a nonlinear analysis is performed and, if at any point the stresses violate the strength limits of the material, fractures occur or grow. In the numerical model, that effect is translated as plastic strain. The coupling between the mechanical and the fluid flow analysis is done by relating the increase in plastic strain with an increase in the permeability on the regions where fracking is present. In other words, the anisotropic
The permeability tensor of the jointed rock mass is updated with a nonlinear relation to the anisotropic plastic strain tensor. This relationship is an essential link between both analyses, and must be calibrated in real world simulations.

4 TEST CASES

On this paper, two hypothetical models will be presented with the aim of showing how the method works, what type of post processing is possible, as well as how the results can actually be seen. Much skepticism surrounds non-classic approaches of fracture propagation, and part of the goal of these models is to shed some light upon one of the different approaches to model fracture propagation.

In a more theoretic point of view, the two models will also be used for a comparison between domains with different material properties. Both domains are exactly the same in terms of geometry, material layer thickness and boundary conditions. The difference lies in the material properties of the individual layers, and such a difference will be evident in the behavior of the fracking during the pumping process.

4.1 Geometry

The geometry is the same for both models. The domain is composed of 5 layers of materials, two of which are modelled as sandstone and the other three are modeled as shale. The layer thicknesses and depths are shown in Table 1.

<table>
<thead>
<tr>
<th>Layer Number (top)</th>
<th>Depth (ft)</th>
<th>Layer Number (bottom)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Layer 1 Top</td>
<td>3320.0</td>
<td>-</td>
</tr>
<tr>
<td>Layer 2 Top</td>
<td>3640.0</td>
<td>Layer 1 Bottom</td>
</tr>
<tr>
<td>Layer 3 Top</td>
<td>4600.0</td>
<td>Layer 2 Bottom</td>
</tr>
<tr>
<td>Layer 4 Top</td>
<td>4920.0</td>
<td>Layer 3 Bottom</td>
</tr>
<tr>
<td>Layer 5 Top</td>
<td>5880.0</td>
<td>Layer 4 Bottom</td>
</tr>
<tr>
<td>-</td>
<td>6200.0</td>
<td>Layer 5 Bottom</td>
</tr>
</tbody>
</table>

The extremities of the domain are modelled far from the pumping point, and form a box with dimensions of 1500 [ft] x 1500 [ft] x 2280 [ft], as can be seen in Figure 1.
4.2 Boundary Conditions and Loads

To set up the initial conditions of the reservoir, pore pressure gradients and overburden stress gradients were defined. For the pore pressure gradient, fresh water with zero salinity was used, generating a gradient of 0.43 psi/ft.

The initial vertical stress was defined by using an overburden pressure gradient of 1.1 psi/ft, which is a mean value of the overburden gradient of shale and low porosity sandstone. Minimum and maximum horizontal stresses were in the same directions on both models.

The nodes are constrained in all degrees of freedom in the boundaries of the domain. Considering that the pumping occurs exactly in the center of the reservoir, rigid boundary effects do not occur in the fracking regions.

On these models, the pumping was applied by prescribing the slurry rate (volume of water injected per unit volume of aquifer per unit time) during the analysis – another option would have been to prescribe the bottom hole pressure (BHP). Only one stage was defined on both examples; the slurry rate was ramped from 0 to 35 ft³/s on the first two minutes of pumping, and was kept at that value until the end of the analysis, 500 minutes later.

It should be noted that the pumping time was defined based on a long enough time to see fracture propagation along the domain and should not be interpreted as a specific pumping time for common hydraulic fracturing applications on real case scenarios.
4.3 Case 1 – Vertical Fracture Growth with Fracture Propagation into Adjacent Layers

- Sandstone layers modeled without joints
  - Fracking continues through intact failure
- Shale middle layer with joints
  - Failure should occur mainly on the vertical joints
- Sandstone intact strength is greater than Shale joint strength
  - A pressure build up should occur before the fractures propagate to the adjacent layers

Despite reaching the extremities of the layer, pumping continues, and the pressure build up results in fracture propagation.

As said before, five layers were modelled in each analysis. Due to the confidentiality of the information, material properties will not be explicitly mentioned. Comparative values between both cases, however, will be mentioned where there are differences. A general overview of the model can be seen in Figure 2.

The first and last layers were modelled as shale, behaving as hard fracture barriers (linear elastic material), to prevent any eventual fluid loss of the domain. The second and fourth layer were modelled as sandstone, working as Soft fracture barriers (nonlinear with intact failure and bedding plane failure) and the shale middle layer was modelled as jointed rock, nonlinear with three joints: bedding plane, and two other joints.

Properties such as cohesion, friction angle and strengths were defined for all joints and for intact rock, as well as their residual values after failure.

Pumping was initiated at the only stage on the perforation, a horizontal stage in the middle of the third layer, respecting the values mentioned above on the boundary conditions topic.
4.4 Case 2 – Vertical Fracture Growth with Fracture Propagation stopping at Adjacent Layers

- Sandstone layers modeled without joints
  - Fracking stops at layer interface
- Shale middle layer with joints
  - Failure should occur mainly on one the vertical joints and the bedding plane
  - Reduced friction angle in shale creates a tendency of bedding plane failure
- Sandstone intact strength is greater than Shale joint strength
  - Lower bedding plane joint strength should keep the fracking in the shale layer

Figure 3 - General overview of case 2.

In terms of layer configuration and the pumping process, this second case is exactly the same as the first one. The main difference between both models is a reduction in the friction angle and the residual friction angle of the bedding plane joint, both of which were reduced by 25%. Such a reduction implies an easier path for in-plane fracture to occur, resulting in a web of fractures contained in the third layer. A general overview of the model can be seen in Figure 3.

This modification is done with the intention of seeing how sensible fracking is to material modification. In case of a high sensibility, such a result would highlight the necessity of obtaining the most accurate material properties as possible.

5 RESULTS

The results discussed in this topic will be divided in two subtopics, to cover individually each of the analysis done. Comparison between both will be made, emphasizing the resulting differences caused by changes in the material properties.

5.1 Case 1

The first step of each case is to check the validity of the initial conditions. The existence of large plastic strains on an initial state would indicate unrealistic in-situ conditions. In other words, it would indicate an instability in the domain even before the fracking occurs, which would be unlikely in a real world scenario.
Figure 4 - The initial results show no plastic strains, indicating a stable in-situ stress condition.

As can be seen from Figure 4, no plastic deformations were caused by the initial conditions, which means that these in-situ stresses are possible to occur. Figure 5 shows the initial vertical and horizontal stresses, and the horizontal stress discontinuities can be clearly seen on the horizontal stresses, due to the different material properties at the interface.

Figure 5 - Effective stresses along the depth

As pumping starts, there is a pressure build up, and material failure occurs at one of the joints. After a certain time of pumping, fractures reach the interface between the shale and the sandstone layers, but a breach only occurs after a subsequent pressure build up, due to the higher intact strength of the sandstone layers, as can be seen in image Figure 6. In the end,
fracture growth on the second and fourth layers was mainly in the vertical direction, caused by the normal pressure on the interface between layers. On the third layer, however, fractures spread further and in more directions. The final configuration of the fractured domain and the plastic strain history can be seen in Figure 7.

Figure 6 - Pressure build up before fractures propagate to adjacent layers.

Figure 7 - Final configuration after 500 minutes of pumping.
5.2 Case 2

In light of the fact that the initial conditions and the elastic properties of both cases are identical, the initial results of plastic strains and vertical and horizontal stresses are also identical to the ones obtained in the first case, and can be seen in Figure 4 and Figure 5.

At the beginning of the pumping process, the fracking occurs in a very similar way as the first case. However, when the fractures reach the interface between adjacent layers, the lower friction angle of the bedding plane on the shale layer causes a flooding on that layer, preventing the pressure build up necessary to frack the neighboring layers. Almost no plastic activity is seen on layers two and four, showing how sensible the results are to material property modifications. The final configuration of the fractured domain can be seen in Figure 8.

![Figure 8 - Final configuration on case two: the fractures are almost exclusively on layer 3 due to the lower bedding plane strength on the middle layer.](image)

6 FINAL REMARKS

Much can be interpreted from the simulation of these two test cases. First of all, in terms of efficiency, both cases were run using only two cores and were finished in a matter of a few hours. For a 3D domain of a whole reservoir with transient fracture propagation, that is a very
reasonable time. On a physical point of view, one can notice that a 25% decrease in one parameter, in this case, an internal friction angle, can cause enormous differences in terms of results. As such, it becomes clear that minor differences can generate a vast amount of possibilities regarding the behavior of the rock and the fracking; also, it is necessary to be very accurate when defining the material properties of each layer.

Not much attention is being given to the interpretation of the individual results of these test cases. That is due to the fact that they are fictional, and interpretations would have no meaning if there is not experimental data or measurements for comparison and calibration.

On real case scenarios, these simulations would be merely the first step of the whole process. Almost every single parameter of input data has a window of uncertainty - particularly the physical properties. Thus, it is necessary to do an initial calibration process of the model, to find, within the windows of uncertainty, the parameter values that best represent the correct behavior of the reservoir based on experimental data such as ISIPs and micro seismic events. That calibration process involves multiple runs through an optimization process to achieve the best results.

After the input data is calibrated, another optimization process is started. Based on a sensitivity analysis to determine the most important of the many input parameters, a series of new simulations are done varying engineering input parameters to generate a response surface of the possible fracking results. Having that, one can determine what type of engineering decisions should be made depending on the necessities of the company doing the simulations. In other words, the simulations done for this paper are only the tip of the iceberg in terms of potential applications using the current approach.

A final point to mention is that the hydraulic fracturing simulator actually uses a lot more input data than was shown on this paper. Much of the necessary input was not explicitly mentioned, but bearing in mind that both cases were fictional and the post processing was done mainly by means of comparison, it was decided to maintain only the most important of these parameters. Nevertheless, it should be clear that many parameters such as those to characterize horizontal stresses, coupling between the hydraulic and mechanical analysis, meshing, permeability and porosity properties, stage directions and so on are necessary inputs.

7 REFERENCES


COMPUTATIONAL MODELING OF VEHICLE-IRREGULARITY-BRIDGE DYNAMIC INTERACTION BY DAMAGE MECHANICS

THIAGO DE O. ABECHE¹, ROBERTO D. MACHADO², FERNANDO L. M. BEGHETTO³ AND LUIZ A. F. DE SOUZA⁴

¹Mechanical Engineering Graduate Program, Pontifical Catholic University of Parana – PUCPR
Rua Imaculada Conceição, 1155 – CEP: 80215-901, Curitiba, Paraná, Brasil
t.abeche@gmail.com

²Mechanical Engineering Graduate Program, Pontifical Catholic University of Parana – PUCPR
Rua Imaculada Conceição, 1155 – CEP: 80215-901, Curitiba, Paraná, Brasil
roberto.machado@pucpr.br

³Department of Building Construction, Federal Technological University of Parana – UTFPR
Rua Deputado Heitor de Alencar Furtado, 4900 – CEP: 81280-340, Curitiba, Paraná, Brasil
flbeghetto@utfpr.edu.br

⁴Department of Civil Engineering, Federal Technological University of Parana – UTFPR
Rua Marcílio Dias, 635 – CEP: 86812-460, Apucarana, Paraná, Brasil
lasouza@utfpr.edu.br

Key words: Dynamic Interaction, Damage Mechanics, Finite Element Method, Computational Modeling.

Abstract. This work deals with the dynamic interaction vehicle-irregularity-bridge through the Finite Element Method, considering the stiffness loss of the bridge by damage mechanics. The irregularities of the track dynamically excite the vehicle, which in turn provokes additional vibrations in the structure of the bridge, besides those produced by their own movement. This condition tends to increase the responses in terms of displacement and stress especially in the resonance conditions. The approach developed in this work treats this phenomenon uncoupled. The track irregularities are represented by sinusoidal harmonic functions. In the bridge model are used Euler-Bernoulli beam elements, with Hermite cubic interpolation functions. The structural damping is defined by the Rayleigh method. The equations of motion are obtained by dynamic equilibrium and numerically integrated in time using the Newmark method. The dynamic response of the bridge is affected by the deterioration degree. In this sense, this work aims to evaluate the dynamic effects produced in a structural model on which the degree of damage is altered over time. This proposal aims to contribute to the study of the health monitoring of bridges structures.

1 INTRODUCTION

The increase in road and railroad transport cargo, especially in Brazil, with increased
intensity values of loads on the roads, has been producing degradation in many bridges throughout the countries. This is an issue of paramount importance, as is related to the structural health of bridges and with the matter of the conservation of roads and railway structures.

In order to contribute with this theme, are being studied the problems of dynamic interaction between vehicle and bridge considering the track irregularities. There are numerous references in the literature dealing with this topic and the modeling of railway vehicle-bridge systems considering track irregularities has been studied by researchers around the world in the last thirty years.

The new dynamic analysis methods for railway bridges are proposed by Goicolea et al. (2002). In this work the impact factor, the dynamic analysis with moving load, analytical methods, finite element method, dynamic analysis with vehicle-structure interaction and models based on harmonic series are presented.

The interaction problem is developed by Cheng et al. (2001). That work presents a bridgetrack-vehicle element for two dimensional analyses. The train is modeled as a series of two degrees of freedom and mass-spring-damper systems. The upper and lower beam elements represent the rails and the bridge deck, respectively. These elements are interconnected by a series of springs and dampers, which represent the sleepers and ballast properties.

Many problems of material degradation are studied with static, iterative and incremental solutions. This study incorporates the damage mechanics in dynamic analysis, turning into a nonlinear dynamic problem. Jacob and Ebecken (1994) describe an optimized computational implementation for nonlinear structural dynamic analysis programs.

2 MATHEMATICAL MODEL VEHICLE-IRREGULARITY-BRIDGE

The model with 1 degree of freedom consists of a non-suspended mass $m_1$, simulating the vehicle wheel, a suspended vehicle mass $m_2$, a spring with stiffness coefficient $k$ and a damper with damping coefficient $c$.

![Simple vehicle model with 1 DOF](image)

When crossing a bridge with speed $v$, the vehicle is subjected to the effects of track irregularities, represented by $y(t)$. This is similar to a case of base excitation and a harmonic base excitation $y(t)$ is here adopted. Assuming a linear elastic behavior for the spring and a linear viscous damping for the damper throughout the analysis, the elastic force acting on the
spring and the damping force are

\[ f_M = k[u(t) - y(t)] \]  

\[ f_d = c[\dot{u}(t) - \dot{y}(t)] \]  

Thus, the governing equation of motion of the problem can be written as

\[ m_2\ddot{u}(t) + c[\dot{u}(t) - \dot{y}(t)] + k[u(t) - y(t)] = 0 \]  

Considering \( v \) as the vehicle speed, \( l \), the wavelength of the irregularities, \( t \) the time and \( A \) a constant, the irregularities functions can be expressed as

\[ y = A\sin\left(\frac{2\pi v}{l}t\right) \]  

Deriving (4) with respect to time, we have

\[ \dot{y}(t) = \frac{2\pi Av}{l}\cos\left(\frac{2\pi v}{l}t\right) \]  

By replacing (4) and (5) in (3), one obtains

\[ m_2\ddot{u}(t) + c\dot{u}(t) + ku(t) = c\left[\frac{2\pi Av}{l}\cos\left(\frac{2\pi v}{l}t\right)\right] + k\left[A\sin\left(\frac{2\pi v}{l}t\right)\right] \]  

The force produced by the base excitation is defined by

\[ F_{EB} = (m_1 + m_2)g + c[\dot{u}(t) - \dot{y}(t)] + k[u(t) - y(t)] \]  

where \( g \) is the acceleration of gravity.

As can be seen in classical literature (Chopra, 1995; Bathe, 1996; Kwon and Bang, 2000), for the Euler-Bernoulli beam finite element, the elementary matrices of stiffness and mass are given respectively by

\[ [K_e] = \frac{E I}{L^3} \begin{bmatrix} 12 & 6L & -12 & 6L \\ 6L & 4L^2 & -6L & 2L^2 \\ -12 & -6L & 12 & -6L \\ 6L & 2L^2 & -6L & 4L^2 \end{bmatrix} \]  

\[ [M_e] = \frac{E I}{420} \begin{bmatrix} 156 & 22L & 54 & -13L \\ 22L & 4L^2 & 13L & -3L^2 \\ 54 & 13L & 156 & -22L \\ -13L & -3L^2 & -22L & 4L^2 \end{bmatrix} \]  

In (8) and (9) equations, the physical parameters \( E \) and \( \rho \) are, respectively, the elastic modulus and the linear mass density of the element, whereas geometric parameters \( A \), \( I \) and \( L \), represent respectively the cross-sectional area, the moment of inertia and the length of the
element.
The equation of dynamic forces on each element is defined by
\[
\{F_e(t)\} = \int_0^L q(x,t)[H]^T \, dx
\]
(10)
where \([H]\) contains the cubic Hermite interpolation functions, like
\[
[H] = \begin{bmatrix}
1 - 3(x/L)^2 + 2(x/L)^3 \\
x[1 - 2(x/L) + (x/L)^3] \\
3(x/L)^2 - 2(x/L)^3 \\
x(x/L)^3 - (x/L)
\end{bmatrix}
\]
(11)
where \(x\) is the abscissa in which are located the efforts in each element, and \(L\) the element length. Substituting (7) in (10) and integrating, we obtain
\[
\{F_e(t)\} = -F_{EB}(t)[H]
\]
(12)
wherein the negative sign is assigned due to the convention system defined previously. The abscissa \(x\) is set as function of time \(t\) and velocity \(v\), therefore \([H]\) becomes
\[
[H^*] = \begin{bmatrix}
1 - 3(vt/L)^2 + 2(vt/L)^3 \\
v[t^2(2vt/L) + (vt/L)] \\
3(vt/L)^2 - 2(vt/L)^3 \\
v[t^2(vt/L)^2 - (vt/L)]
\end{bmatrix}
\]
(13)
Rewriting the (12) using the (13), we obtain
\[
\{F_e(t)\} = -F_{EB}(t)[H^*]
\]
(14)
The global equation of motion of the undamaged bridge is written as
\[
[M_B]\ddot{U}_B + [C_B]\dot{U}_B + [K_B]U_B = \{F_B(t)\}
\]
(15)
where \([M_B]\), \([C_B]\) and \([K_B]\) are the global matrices of mass, damping and stiffness, \(\{U_B\}\), \(\dot{U}_B\) and \(\ddot{U}_B\) are the global vectors of displacement, damping and acceleration, and \(\{F_B(t)\}\) is the external force vector. Here, the Newmark method is utilized with average acceleration to time integration of Eq. (15), and the system is solved by Gaussian elimination method.

3 MAZARS DAMAGE CONSTITUTIVE MODEL

In this work, it is considered that the bridge is made of reinforced concrete and, when passing over the bridge, the vehicle produces some damage that will influence the dynamic response of the bridge.

The Mazars Damage Model (1984) is used, which is based on some experimental evidences observed in uniaxial experiment in concrete, having the fundamental hypotheses:

- the damage is represented by a scalar variable \(D\) \((0 \leq D \leq 1)\) whose evolution occurs
when a reference value for the ‘equivalent stretching’ is exceeded;
– locally the damage comes from the existence of stretching deformations;
– it’s considered, that the damage is isotropic, although experimental tests show that the
damage leads, in general, to an anisotropy of concrete (which may initially be considered as
isotropic); and
– the damaged concrete behaves as an elastic medium although the permanent deformation
evidenced experimentally in a situation of unloading is neglected.

The square of the equivalent strain is equal to the sum of the squares of the main
components of the positive principal strain
\[
\tilde{\varepsilon}^2 = \sum_{i} \left\langle \varepsilon_i^2 \right\rangle_+ \quad (i = 1, 2, 3, ..., \infty)
\]  
(16)

The extension state is locally characterized by a stretching or an equivalent strain,
expressed as (Pituba, 1998)
\[
\tilde{\varepsilon} = \sqrt{\left\langle \varepsilon_1^2 \right\rangle_+ + \left\langle \varepsilon_2^2 \right\rangle_+ + \left\langle \varepsilon_3^2 \right\rangle_+ + ... + \left\langle \varepsilon_n^2 \right\rangle_+} \quad (i = 1, 2, 3, ..., \infty)
\]  
(17)

where \(\varepsilon_i\) are the principal strain components and \(\left\langle \varepsilon_i \right\rangle_+\) its positive parts defined by
\[
\left\langle \varepsilon_i \right\rangle_+ = \frac{1}{2} \left[ \varepsilon_i + \left| \varepsilon_i \right| \right]
\]  
(18)

The damage starts when the equivalent strain \(\tilde{\varepsilon}\) reaches a value of the reference strain \(\varepsilon_{d0}\)
determined in uniaxial traction tests in correspondence to the maximum stress, as the
following figure

![Figure 2: Stress-strain diagram for Mazars damage model](image)

The constitutive relation, for the particular case of unidimensional stress state, is given by
(Tiago et al., 2002)
\[
\sigma = (1 - D(\varepsilon))E_{e0}\varepsilon
\]  
(19)
in which \(E_{e0}\) being the initial elastic modulus, that is, of the non-damaged material.
The basic damage variables are given by
\[
D_T(\tilde{\varepsilon}) = 1 - \frac{\varepsilon_d(1 - A_T)}{\tilde{\varepsilon}} - \frac{A_T}{e^{\varepsilon_d(1 - \varepsilon_{d0})}}
\]

\[
D_C(\tilde{\varepsilon}) = 1 - \frac{\varepsilon_d(1 - A_C)}{\tilde{\varepsilon}} - \frac{A_C}{e^{\varepsilon_d(1 - \varepsilon_{d0})}}
\]

where \(A_T, B_T, A_C\) and \(B_C\) are the characteristic parameters of the material in uniaxial traction and uniaxial compression, respectively, \(\tilde{\varepsilon}\) the equivalent strain below which no damage occurs and \(\varepsilon_{d0}\) the parameter of the limit elastic deformation. The subscripts \(T\) and \(C\) refer to traction and compression, respectively. Therefore, if the equivalent strain is lesser then the reference strain \((\tilde{\varepsilon} < \varepsilon_{d0})\), then there is no damage at all \((D = 0)\).

Through a linear combination of the basic variables of damage, \(D_T\) and \(D_C\), with the combination coefficients, \(\alpha_T\) and \(\alpha_C\), the damage variable \(D\) is given by

\[
D(\varepsilon) = \alpha_T D_T + \alpha_C D_C, \quad \alpha_T + \alpha_C = 1
\]

where the values of the coefficients \(\alpha_T\) and \(\alpha_C\) contained in the closed interval \([0,1]\), and represent the contribution of requests to traction and to compression for the local state extension, respectively (Pituba and Proença, 2005).

Mazars (1984) proposed the following ranges of variation for the parameters \(A_T, B_T, A_C\) and \(B_C\), obtained from the calibration with experimental results (Pituba, 1998)

\[
0.7 \leq A_T \leq 1 \quad 10^4 \leq B_T \leq 10^5 \quad 10^{-5} \leq \varepsilon_{d0} \leq 10^{-4}
\]

\[
1 \leq A_C \leq 1.5 \quad 10^3 \leq B_C \leq 2.10^3
\]

Figure 3: Uniaxial response of the Mazars model: (a) traction, (b) compression (adapted from Pituba, 1998)

4 STRUCTURAL STEEL CONSTITUTIVE MODEL

It is used a uniaxial model to describe the behavior of the steel reinforcement, since in reinforced concrete structures the steel bars resist fundamentally the axial forces. The structural steel is represented as an elastoplastic material having the same behavior in traction and in compression. The representation is given by a bilinear stress-strain diagram (Pasa, 2007).
Thus, the stress acting on the structural steel is determined by (TIAGO et al., 2002)

\[
\sigma = \begin{cases} 
E_s \varepsilon, & \text{if } -\varepsilon_{sy} \leq \varepsilon \leq \varepsilon_{sy} \\
E_{sy} \varepsilon, & \text{otherwise}
\end{cases}
\]  

\((23)\)

where \(E_s\) is the initial elastic modulus of the structural steel, \(\varepsilon_{sy}\) is the yield extension and \(E_{sy}\) is the longitudinal elastic modulus after the yield of the steel defined by

\[
E_{sy} = k_s E_s
\]  

\((24)\)

wherein \(k_s\) is the relation between the longitudinal elastic modulus after the yield of the steel \(E_{sy}\) and the longitudinal elastic modulus of the steel \(E_s\).

5 EQUIVALENT STIFFNESS

Since the continuum damage mechanics is considered dynamically, it is necessary to evaluate the damage in each section of the bridge, i.e., for each iteration. In this sense, the cross section of the beam is divided into \(n\) layers, as laminated composite beams, as shown in Figure 5, to be possible determine the equivalent stiffness.
For a particular case of symmetric laminated composite beam with \( b \) width, the equivalent bending stiffness \( EI_{\text{eqv}} \) is determined by

\[
EI_{\text{eqv}} = \frac{1}{3} b \sum_{i=1}^{n} E_i \left( y_i^3 - y_{i-1}^3 \right)
\]

where \( n \) is the number of layers, \( b \) is the width of the rectangular cross section, \( E_i \) is the elastic modulus of the \( i^{th} \) layer, in the case \( E_c \) for the concrete layers and \( E_s \) for the steel layers, \( y_i \) and \( y_{i-1} \) are the \( y \) coordinate values of the division points of the \( i^{th} \) layer which subtracted result in the height of the layer.

In obtaining the elementary internal force vector, the equivalent bending stiffness is determined for each Gauss point in the numerical integration by using the Gaussian quadrature method.

In the calculation process of the equivalent stiffness, when the materials, concrete or steel, present nonlinear behavior, the position of the neutral axis varies and is recalculated at each numerical iteration according to the deterioration of any layer.

The deformation state, one-dimensional or three-dimensional, of a damaged material is obtained from the intact material behavior law where the normal stress is replaced by the effective stress, in order to obtain a coherent model with the hypothesis of continuum medium for the equivalent strain (Lemaitre and Chaboche, 1985).

This way, assuming a linear elastic behavior for the intact portion of the medium, in the unidimensional case the elastic strain is given, according to Pituba (1998), by

\[
\varepsilon_c = \frac{\sigma}{E} = \frac{\sigma}{(1-D)E}
\]

where \( E \) is the elastic modulus of the intact material. The Young’s modulus of the damaged material, \( \tilde{E} \), for a continuum medium of equivalent response to the imperfect material, is defined by

\[
\tilde{E} = (1-D)E
\]

This relation allows to indirectly determining the damage variable for elastic materials from measures of the elastic modulus for a test with loading and unloading cycles.

Thus, in each iteration for each time step, it analyze if the strains exceed the criterion reference strain and, if so, damage occurs and the elastic modulus is updated.

The problem in question first calculates the dynamic responses of displacement, velocity and acceleration with the intact bridge, without damage, by solving the Eq. (15). On other hand, when damage occurs the forces are no longer linearly dependent of the displacements, causing the previously fixed stiffness matrix to become instantaneous, characterizing nonlinearity, as shown below

\[
[K_b] = [K_b(U_b)]
\]

Therefore, the global equation of motion of the damaged bridge is written as (Jacob and Ebecken, 1994)
Here, the above nonlinear equation (29) of motion is solved in incremental and iterative way by the combination of the modified Newton-Raphson iterative technique with the implicit time integration operator of the Newmark method.

6 NUMERICAL ANALYSIS

The numerical analysis shown below is intended to compare the linear and nonlinear dynamic analysis in terms of displacements, velocities and accelerations for both the intact and the damaged bridge for the following input data. These models and procedures are computationally implemented in a MATLAB® code.

A simply supported beam has 20 m length, a rectangular cross section of 0.4 x 3.9494 m, a mass per unit length of 34,088 kgf/m, the first natural frequency of vibration of 44.75 rad/s, a Young’s modulus of 29.43 GPa for the concrete and 210 GPa for the structural steel, a Poisson’s ratio of 0.2 for the concrete and 0.2 for the structural steel. The length and the amplitude of the sine wave of the irregularities are 1 m and 0.005 m, respectively. To determine the equivalent stiffness, the cross-section of the beam was divided into 60 layers. The damping ratio used was 0.025. In the discretization by finite elements were used 20 beam elements with 2 nodes per element and 2 degrees of freedom per node. In the bilinear elastoplastic model for the structural steel were adopted $k_s = 0.85$. The reinforcement ratio used was 2%.

The vehicle is crossing the bridge with 50 km/h of speed, a stiffness coefficient of 9.120 kN/m, a damping coefficient of 96 kNs/m, a non-suspended mass of 44 tf and a suspended mass of 176 tf. It is noted that vehicle masses are exaggerated. Nevertheless, this condition is necessary to achieve the desired level of damage by the authors in the dynamic responses shown below.

The Newmark parameters chosen were 0.5 and 0.25 with 10,000 time steps. The Mazars damage parameters adopted were: $A_f = 0.995$, $B_f = 30.000$, $A_c = 1.2$, $B_c = 1.050$, and $\varepsilon_{d0} = 10^{-4}$.

The graphics generated in the following analysis were elaborated with a point $P$ located in the middle of the beam structure.

Below are shown the graphics of the comparison between the dynamic responses of displacement, velocity and acceleration, respectively, for the linear analysis of the intact bridge and for the nonlinear analysis of the damaged bridge, nonlinear.
It is noticed that the dynamics displacements along the damaged bridge are greater than the dynamic displacements without the presence of the damage. This happens because the loss of intact resistive area and consequent stiffness loss cause greater effective stresses. The track irregularities associated with the vehicle vibrations caused greater oscillations in dynamic displacement responses of the damaged bridge.

Figure 6: Comparison between dynamic responses of displacement (linear and nonlinear) in the middle of the beam structure

Figure 7: Comparison between dynamic responses of velocity (linear and nonlinear) in the middle of the beam structure

It is possible to observe that the speeds with damage influence have greater variations
within the same time interval. The time response is coincident because the event is coincident, but occurs amplification of the oscillations. The beating phenomenon is noted in nonlinear dynamic response indicating the presence of resonance. Although not intentionally emphasized in Figure 7, it is possible to observe the damping of the system. Greater amplification of the oscillations can be noted in nonlinear response.

Figure 8: Comparison between dynamic responses of acceleration (linear and nonlinear) in the middle of the beam structure

Observing the graphic above, it is clearly noted the beating phenomenon in nonlinear dynamic response of acceleration, characterized by a rapid oscillation with a low amplitude variation due to superposition of waves with same direction, same amplitudes and nearby frequencies (Inman, 1996). The values of the dynamic responses of acceleration, linear and nonlinear, vary between positive and negative values. The accelerations with damage influence have greater variations within the same time interval. There is also greater amplification of the oscillations in nonlinear response.

7 CONCLUSIONS

In the bridge analysis, beat motion occurs in nonlinear dynamic responses indicating resonance conditions. The study showed that before excessive loads that produces damage on the structure, the dynamic responses becomes completely modified. The dynamic responses of displacement were greater with the presence of the damage. There were major variations in nonlinear dynamic responses of velocity and acceleration within the same time interval. The dynamic responses of velocity and accelerations ranged between positive and negative values. There was greater amplification of the oscillations in all nonlinear responses.

The present study has been considered through the application of a high intensity load. In practical terms and in addition to the conclusions related to this work, it is appropriate mention that this program could be used, in future works, to include damage by material fatigue over time to the bridge and repetitive loads in unusual situations, thereby contributing to the
monitoring of these structures.

Only the displacement, velocity and acceleration dynamic responses are included in this paper. However, full results which include the effects of dynamic amplification factor by varying speeds, for both intact and damaged bridge, and the damage degree of the structure have been performed and will be reported in the master’s thesis of the first author.

Acknowledgements

The first author would like to gratefully and sincerely thank the other authors and all the teachers of PPGEM/PUCPR, also acknowledge the financial support provided by CAPES.

REFERENCES

PROTECTIVE STRUCTURES – A CHALLENGE TO COMPUTATIONAL ENGINEERING AND TESTING

Norbert Gebbeken, Tim Bermbach

University of the Bundeswehr Munich, Institute of Engineering Mechanics and Structural Mechanics
Werner-Heisenberg-Weg 39, 85577 Neubiberg, Germany
e-mail: norbert.gebbeken@unibw.de, tim.bermbach@unibw.de
webpage: http://www.unibw.de/baustatik

Keywords: laminated glass, air blast load, material modelling, secondary debris, shock tube experiments

Abstract. Contemporary architecture highlights building envelopes made of glass. Facade technology has become a new field of technology in the last years. However these facades are vulnerable when hurricanes, tornados or blast have to be taken into account. Therefore, protective glazing is an important field of research and development. Surprisingly, there are papers that are inconsistent in the research results. This is amongst others due to the fact that no guidelines for testing exist. Another problem is the interpretation of test results. The present paper will discuss the state of the art in design as well as research and development. Own projects and studies will be presented.

1 INTRODUCTION

The rising use of glass as building envelope requires considerations concerning the resistance against extraordinary loads. The article focusses exemplarily on laminated glass exposed to an air blast wave. Firstly, the idealization of this high dynamic load and the general response of the composite are discussed. When a numerical model is assembled chosen material laws must be able to reproduce characteristics leading to specific observations. The components of laminated glass (glass, interlayer) show peculiarities in terms of material behavior. Furthermore the user has to select an approach corresponding to the objective of the analysis. The scope of the task defines accuracy, the level of detail and consequently numerical methods needed for accomplishment. Finally, attention is turned on the adhesive bond among composite layers which represents an important factor in regard to initial and residual load carrying capacity. Conclusions from planned shock tube testing shall be applied on computational contact modelling.

2 STRUCTURES LOADED BY AIR BLAST WAVES

2.1 Air Blast Wave Modelling

Air blast waves result from a rapid release of energy. The magnitude of the overpressure of an air blast wave arriving at a specific location depends on the distance and on the size of the charge. An idealized free-air blast form of a pressure-time function at a certain distance from the explosive is shown in Figure 1.
The main characteristics of a free-field air blast wave are the following:

- The shock wave of the air blast reaches the point under consideration at the *arrival time* $t_a$,
- The pressure attains its maximum (*peak overpressure* $p_{max}$) very fast (extremely short rise time). Then it starts decreasing until it reaches the *reference pressure* $p_0$ (in most cases the atmospheric pressure),
- The *duration of the positive phase* $t_d$ is the time taken to reach this reference pressure. After this point the pressure drops below the reference pressure until it reaches the *maximum negative pressure* $p_{min}$. The *duration of the negative phase* is denoted as $t_n$,
- The overpressure impulse (positive impulse) is the integral of the overpressure curve over the positive phase $t_d$.

The idealized form of the pressure wave can be significantly altered by the morphology of the medium encountered along its propagation. For instance, the incident peak pressure can be magnified up to 14 times if the wave is reflected at a rigid obstacle. The reflection effects depend on the geometry, material, size, and angle of incidence (see Gebbeken\(^1\) for further information). The situation is much more complicated if there are several reflection boundaries as it happens in urban environments.

A widely used way of describing the shape of the air blast pulse is the so-called modified Friedlander equation, which proposes a function for the positive phase of the air blast wave. The pressure $p$ at time $t$ can be calculated by

$$ p(t) = p_0 + p_{max} \left( 1 - \frac{t}{t_d} \right)^{-b_t} $$

All parameters of Equation (1) can be taken from several diagrams and equations (e.g. Kingery\(^2\)). The parameters are available for incident and reflected waves for spherical free-field as well as for hemispherical charge conditions.
The negative phase (peak pressure and duration) can be described with a diagram proposed by Drake\textsuperscript{3}, obtained from experimental data. A bilinear function for the negative part of the air blast wave is used in this work.

The decay of the pressure-time function can be adjusted with the form parameter b of Equation (1). Several diagrams from the literature can be used for this parameter (e.g. Baker\textsuperscript{4}).

### 2.2 Laminated Glass Failure due to Air Blast Load

Laminated glass is built of two or more float, annealed or toughened glass panels combined with one or more polymeric interlayers. Secondary debris of glass is the most probable cause for injuries in case of an explosion event. The aim of laminated glass is the reduction of ejected glass splinters in case of failure by its adhesive bond. According to Larcher\textsuperscript{5}, the failure of a laminated glass pane can be subdivided in five phases as shown in Figure 2:

![Figure 2: Schematic failure of laminated glass\textsuperscript{5}](image)

(1) Elastic behavior, intact glass plies,
(2) The first glass ply is broken, the other is still intact,
(3) The second glass ply fails; the interlayer behaves elastically,
(4) The interlayer acts as membrane,
(5) The interlayer fails by reaching its failure strength or by cutting of glass fragments.

Phase (1) and the first fragile failure of glass can be modelled with either analytical or numerical methods. However simulation of phases (2) to (5) is more complex.

Besides the global behavior of the composite, local effects have to be considered as well. During phase (2) bending stresses are relocated to the intact cross-section. Even in phase (3) the composite bears bending as compressive stresses are transmitted among remaining glass fragments while the interlayer is acting as tie bar. Analogous to tension stiffening of reinforced concrete, these fragments cause a tension stiffening effect in phase (4).
3 MATERIAL MODELLING OF LAMINATED GLASS COMPONENTS

Choosing or implementing a material model for laminated glass the specific characteristics of its components have to be understood. Depending on desired accuracy of the model or investigated event these characteristics should be taken into account.

3.1 Material Law Considerations for Glass

Glass behaves linear elastic until brittle failure. The designation of tensile strength differs in a theoretical and a practical value. Theoretical tensile strength of glass is determined by the strength of chemical bonds on molecular level and can be calculated with up to 30.000 MN/m². In common testing glass never reaches this theoretical value as micro cracks on the surface of glass plies are pervasive and initiate failure by crack propagation. This leads to a reduced practical allowed tensile strength of 75 to 150 MN/m².

Therefore the practical tensile strength of glass can be interpreted less as material constant but rather as measure of surface quality. The distribution of micro cracks and consequently the tensile strength on a surface of a glass ply can be described by a Weibull distribution, which corresponds to the figurative failure of the weakest link within a chain. Against this background failure may be initiated from less loaded locations.

As failure criterion of brittle materials the maximum principal normal stress theory by Rankine is widely spread. However the tensile strength of glass is dependent on a multiaxial stress state, which has been discovered in material testing by Siebert.\textsuperscript{6} Regarding the tensile strength an increase of 20 % in comparison to uniaxial tests could be measured in multi axial testing.

Beside material constants further propagation after generation of an initial crack is dependent on the factor of stress intensity, which is characterised by the original crack geometry and the local mode of stress distribution in immediate vicinity of the crack. With regard to the material dependency temperature, air humidity and chemical composition of the glass play a role.

Thermal or chemical post-tensioning increases the mechanical resistance of the glass surface prior cracking. In terms of elastic fracture analysis the value of post-tensioning resistance is irrelevant, though its magnitude in manner of released energy affects the crack pattern: after failure toughened glass shows a smooth crack pattern whereas a coarse fragmentation is typical for annealed glass.

Loads can be transmitted across cracks by friction and compression in-crack-plane between glass fragments. Hence the compressive strength has to be brought into focus when modelling laminated glass in a failure state. Material testing by Fink\textsuperscript{7} shows an increase of compressive strength with augmenting strain rate.

3.2 Material Law Considerations for polymeric Interlayers

Most commonly used interlayer materials are polyvinyl butyral (PVB), thermoplastic polyurethane (TPU), ethylene acetate (EVA) and ionoplast polymers (IPP). Regarding the specific temperature range being relevant for usual civil engineering applications, the properties of these basic materials vary severely. Furthermore properties of one and the same basic material deviate from each other as a result of different additives used in the production process.

Widening the view on the temperature range common grounds of material behavior among different polymers are found. The transition temperature describes the start of converting a polymer from a glassy into a viscous state when increasing the temperature. For example IPP owns a transition temperature of about 55 °C, PVB of about 12 to 20 °C.
In addition the material properties of polymers show a clear strain rate dependency. Regarding the testing results of Kott & Vogel, an increase of the strain rate leads to a glassy behavior although testing temperature is located above transition temperature. This phenomenon could be interpreted as a reduction of the transition temperature.

Comparing stress-strain-curves of uniaxial tensile tests the rheological behavior naturally differs below and above the transition temperature. Above transition temperature a sloping curve characterized by a secant modulus describes the viscoelastic stress-strain-trend. Below transition temperature a linear elastic stress-strain-trend characterized by the Young’s modulus is followed by a sloping curve characterized by a secant modulus.

Reliable and broad testing of polymers with variable strain rate and temperature cannot be found in literature. Concerning the second range of the stress-strain-trend below transition temperature the few testing results available differ in their evaluation. Morison mentions explicitly the recuperation of any deformations and thus assumes full elastic behavior of the material. On the contrary, Kott & Vogel, Iwasaki and Bennison observed plastic deformations. However, plastic behavior can only be characterized if, and only if, the sample has been unloaded in order to measure the remaining strain. This is missing in the papers. Thus, the observation of plastic behavior is questionable.

Another important aspect related to material behavior of polymers is weathering, which has been investigated by Ensslen and Sackmann. Hot and dry climate conditions in combination with radiation exposure reduce moisture and plasticiser content of the interlayer and induce brittlement. In an environment of high air humidity absorption of moisture due to hygroscopic behavior can cause a reduction of tensile strength of around 85%.

4 NUMERICAL ASPECTS

4.1 Simulation of Air Blast Waves

There are several methods of numerical modelling that can be used in order to load a structure with an air blast wave. These methods differ in type of elements, number of finite elements used, implicit or explicit calculation, and others. The choice among these methods depends on the scope of the analysis and the available calculation time.

- The solid TNT model describes the mechanical behavior of the explosive with a material law, e.g. the JWL equation. A fine mesh is essential to obtain realistic results. The calculation is therefore very time consuming. If the mesh is not fine enough, the computed pressure and impulse are underestimated,
- Compressed balloon (Larcher). The pressure-time function resulting from a compressed balloon can match fairly well the curve of an air blast wave. The amount of pre-compression can be calibrated using the resulting impulse. The calculation time is shorter than for the solid TNT model,
- Smooth Particle Hydrodynamics (Hiermaier). The geometry of the explosive is discretized as a crowd of particles. Any particle can be taken as supporting point regarding thermodynamic and kinetic variables. Impulse is transmitted among the particles. This method offers advantages of a Lagrangian formulation without mesh dependency,
- Load-time function. The structure is loaded by a load-time function built with the pressure-time function presented in chapter 2.1. The calculation is relatively inexpensive. However, the method cannot represent reflections, shadowing and
channelling. A load-time function should in principle be used when the loaded structure can be considered as rigid. Large displacements of the structure would reduce the reflected pressure load.

4.2 Finite Element Modelling Methods

For describing the laminar composition of safety glass several modelling techniques are available. Due to the short duration of an air blast wave, full bond of the laminated layers can be assumed. A few methods do not consider the laminar composition, so that monolithic shell or solid elements are employed. This approach is in agreement with experimental observations (see Kolling[16]), though there are deficiencies in reflecting crack propagation and composite behavior in a damaged state correctly. If shell and solid elements are taken into account, a variety of possible combinations will be found already. In the following commonly applied methods are presented.

- **Coupled shell elements - single stage failure criterion.** This simple method consists of two coincidently connected shell elements. The elements share the same nodes and the center planes of both elements are superposed. Thickness, density and Young’s modulus of the elements are calculated in such a way that combination of both represents the laminate behaviour before failure. A single shell is representing laminate behavior in a user-defined failure state. For example, entire failure of all glass sheets and corresponding loss of bending stiffness at once could be assumed. The material properties of glass including a failure criterion are attributed to one shell. The other shell represents the polymeric interlayer with sticky glass fragments, which acts as membrane after failure and erosion of the element mentioned firstly. Another failure state like a remaining single glass ply as found in case of a blunt impact could be assumed as well. In consequence shell properties have to be adapted,

- **Coupled shell elements - two-stage failure criterion.** With this method the actual laminar composition is regarded and an accurate bending stiffness of the composite is depicted. Modelling of glass panels on attack side and attack averted side failing independently from each other is possible,

- **Multilayer shell element.** The whole composition is represented by a single shell element. The Gauss evaluation points along shell thickness are assigned to the respective material properties,

- **3D solid elements.** This method requires very fine meshing, particularly in through-thickness direction, which causes a long calculation time depending on the applied integration rule,

- **Finite elements combined with Smooth Particle Hydrodynamics.** If secondary debris on the attack averted side is of particular interest the laminate will be discretized locally in form of particles. This method reproduces large strains and behavior of the material after failure. Coupling of Smoothed Particle Hydrodynamics with finite elements is treated in Sauer[17].

4.3 Simulation of Supports and Drills

Modelling of boundary conditions is crucial in order to receive the correct response of a structure in a simulation. The behavior of the substructure and the numerical capture of local stress concentrations may be overriding and must therefore be considered.

In facade construction glass panels are usually mounted on flexible structures. In case of a blast wave the structure interacts dynamically with the surrounding medium.
Emerging effects have been investigated by Teich. It is found that the reduction of pressure depends on the lowest Eigen frequency of the structure and that the peak pressure is not reduced significantly. Furthermore the material stiffness of spot or linear supports itself may be incorporated in the model. In the majority of cases glass panes are isolated from the support by an elastomeric or thermoplastic layer. Relying on a detailed model the stiffness of this isolation material may be determined to reproduce actual support conditions. Clamping supports pose a challenge to the design against disengagement of the pane. Friction among contact surfaces may be modelled with a specific contact formulation.

Concerning the numerical capture of stress concentrations respective regions like support areas and drills need to be discretized in a higher resolution. A convergence study is essential so that relevant information does not drop away due to a coarse model. The tangential-radial-ratio of element dimensions in the area of discretisation near drills and spot supports should be kept around 1 to avoid numerical instabilities.

5 Shock Tube Experiments

In the last quarter of 2014 shock tube testing of laminated glass panels will be conducted at the facilities of Bundeswehr Technical Center 52 in Oberjettenberg. Test series aim at giving input for modelling the contact among layers numerically. Investigations are focused on the bond between glass and interlayer, the course of damage of the temperature and strain rate dependent composite as well as the assessment of secondary debris hazard on the attack averted side.

5.1 Adhesion among Laminate Layers

The bond among glass and a polymeric interlayer is achieved on molecular level. Keller mentions moisture and residual salts as main factors for a reduction of adhesion which can be the consequences of a poor quality control during the production process. In matters of impact loads a controlled reduced bond between the layers is favoured to gain energy dissipation of the interlayer by local delamination in vicinity of glass cracks.

5.2 Test Set-Up

Tested composite pane is clamped in a frame which is fixed on a rigid substructure. A preliminary assessment is necessary to determine the strength and the length of clamping. On the one hand venting and disengagement of the pane due to large deflections has to be avoided. On the other hand an adjusted stiffness shall reduce the risk of undesired shear failure near the supports.

Digital Image Correlation is used to determine kinetic variables of the pane. A thermal imaging camera records the temperature trend on the pane surface. Additional sensors quantify pressure, air humidity and temperature.

The area on attack averted side is divided into sections against distance and angle to the pane. The sections are covered with material directly intercepting secondary debris. A schematic illustration of the test set-up can be found in Figure 3.
5.3 Execution

All panes tested are equal in their symmetric buildup glass - interlayer - glass but differ in interlayer material, manufacturing technique of the lamination and finishing-treatment of the glass. Furthermore the composite will be tested for two temperatures. Variations are summarized in Table 1.

<table>
<thead>
<tr>
<th>Interlayer and Technique of Lamination</th>
<th>Glass Plies</th>
<th>Temperature of composite</th>
</tr>
</thead>
<tbody>
<tr>
<td>PVB, autoclaved</td>
<td>float</td>
<td>12 °C</td>
</tr>
<tr>
<td>PVB reduced adhesion, autoclaved</td>
<td>toughened</td>
<td>25 °C</td>
</tr>
<tr>
<td>IPP, autoclaved</td>
<td>annealed</td>
<td></td>
</tr>
<tr>
<td>TPU, autoclaved</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EVA, vacuum method</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Variation of parameters in shock tube testing

Incremental increase of the air blast load is appropriate to validate subsequent numerical simulations in different phases of damage. Therefore a configuration is tested several times, so that following responses are attained varying the mass of the explosive:

- Elastic behavior of the composite,
- Failure of glass plies but intact interlayer and
- Failure of glass and rupture of the interlayer.

3.3 Evaluation

Besides measuring signals the investigation of secondary debris and the finite state of the damaged composite is of particular interest. The areas of the surface on attack averted side are assigned to the damage categories

- Intact,
- Cracked glass without ejection of glass splinters,
• Cracked glass with ejection of glass splinters, cohesive failure within glass layer.
• Cracked glass with ejection of glass splinters, adhesive failure at interface glass-interlayer and
• Failure of glass and rupture of interlayer.

Intercepted glass splinters in rearward sections are collected, categorised by size and weighed by specific section and category to determine the kinetic energy related to the areal distribution of debris. Representative and peculiar splinters are examined microscopically at random.

6 CONCLUSIONS

In this paper state of the art modelling of laminated glass under high dynamic loads is presented. Simulation of phenomenological observations requires understanding of specific material characteristics. Broad material testing of polymers with variable strain rate and temperature needs to be undertaken as the polymeric interlayer strongly influences the response of the laminate in the damaged state. Several numerical approaches are available and have to be chosen dependent on the scope of analysis. The bond among composite layers will be explored in own shock tube experiments to capture contact computationally.

REFERENCES


THE ANALYSIS OF SPLICE JOINT BEHAVIOUR IN HEAVILY LOADED TENSILE CHORDS OF WARREN TRUSSES

LILITA OZOLA

Department of Structural Engineering, Latvia University of Agriculture
Akademijas 19, Jelgava, Latvia, LV-3001
E-mail: Lilita.Ozola@llu.lv

Key words: Steel Structures, Warren Trusses, Splice Joints, Modelling.

Abstract. This current study deals with the analysis of heavily loaded splice joint behavior in tensile chord of Warren trusses. Theory of statically indeterminate beam on elastic supports has been applied for the modelling of end-plate behaviour. To gain deeper insight into the mutual behaviour of splice joint elements of tensile chord, some issues are discussed in this current paper: 1) suggestions for modelling of splice joint in tensile chord considering additional stresses induced by bending moment; 2) drawing up the methodology for analysis and numerical examples of end-plate joint design considering elastically deformed tensioned bolts and stiffness on end-plate; 3) determination of more unfavourable values of internal forces in elements depending on truss topology and the design model selected, underlining the link between stiffness properties of a structure and bending moment values expected. A design model is presented for end-plate connection acceptable for analysis by using the common software (MS Excel, MathCAD) for engineering calculations. The end-plate is modelled as a continuous beam on discrete elastically deformed supports (representing tying bolts), and loaded by concentrated forces transferred by flanges and the web of I section correspondingly. It has been concluded that the design models leading to disregarding some portion of stresses or strains in the comparison with real structural behaviour, may be assessed as simplified and insufficient for the design of bearing structures, and always a thorough analysis of stress-strain relationships developing as uppermost when overloading becomes real. Therefore it is necessary to increase the requirements for risk assessment of public building structures, and particularly trade buildings/shopping centres concerning the recent bad experiences of the past year’s accidents.

1 INTRODUCTION

During the last few decades most efforts have been devoted to the elaboration of automatic design tools for structures, and a high level of performance and cost-effectiveness have been achieved. Sometimes a comprehensive design practice entails some risk when the most unfavourable loading situations for the whole structure or stress-strain relationships are not analysed. Tolerance of this risk varies widely among countries and emerges from the legal code system and content, and clarity in definitions of criteria for the acceptability of construction products, and even more from the intelligence of professionals. It is significant when considering the risk to quantify it as part of a conscious decision making process in construction.
Some controversial solutions which are implemented in design practice have raised an issue about a more extensive study of the stress-strain relationships in structural elements and the behaviour of joints specifically. These disputable topics in the truss design area deal with the conformity between design model links and a realistic behaviour phenomena of a system. The overall objective of this article is to point out some specifics of structural composition, to promote comprehensive thinking in design, therefore making it possible to ensure the safety of the structures. The specified method used is the phenomenological approach to the splice joint behaviour problem illustrated by the results of numerical analysis based on a real design solution.

2 STRUCTURAL AND DESIGN MODELS OF HARD LOADED WARREN TRUSS

The use of modern computational technologies enable effective tools for analysis of a wide scope of structural models, and of high complexity. Consequently, the major issue for the structural engineer is the choice of a valid model from several acceptably adequate ones. In any case the results of the analysis involve some uncertainty in the design.

It is routine in practice to treat the general model of the truss representing the web members as a pin joined to the continuous chords, see Fig.1b. In most cases no more static trials have been carried out, and consequently such simplification leads to overlooking some portion of the stresses.

Detailed analysis of different static models proposed for simulation of the behaviour of the bearing structure more correctly becomes significant for heavily loaded trusses, and particularly when the junction has been put into a tensioned chord element due to the requirements of transportation (Fig. 1a)

The opinion suggested by this current paper is illustrated by a practical example of the Warren truss design with an effective span of 24 m (≈ 78.7 ft) under vertical load (design value \(q\) = 46.1 kN/m (≈ 3.2 kip/ft) applied directly to the upper chord through concrete slabs.

Welded connections between the chord members and diagonals provide rigidity of a joint capable of preventing any rotation of the end sections of members when the structure behaves under load. This model (Fig.1c) is applicable for the determination of a full scale internal force assembly: bending moments, shear forces which must be taken into consideration together with the axial forces. Particularly significant are the differences of axial force values for tensioned diagonals at the supports depending on the calculation model used. Using the conventional model some portion of the axial force is ignored. More impressive results are shown in Table 1. Sections of support diagonal and bottom chord of heavy rolled profile HEA 260 (section area \(A\) = 8680 mm\(^2\) (≈13.45 in\(^2\)), section modulus about strong axis \(W_y\) = 836400 mm\(^3\) ≈51.04 in\(^3\)) show significantly distinctive results in loading prognosis, and more unfavourable ones should be taken into account in design. Besides it is important to note that bending moment produces the additional portion of normal stress perpendicular to the throat of the filled weld at the joint.

Another phenomenon, usually not considered in practice is the option of putting in or removing the middle vertical bar called zero one in static meaning, see joint C (Fig. 1c). Nevertheless it is a significant element when considering the bending moment in the bottom chord sections. It is a demonstrative static trial of deformable contour. A contour without a middle bar is characterized by significant differences (in structural meaning) of the middle
point displacements of upper and bottom chord: the vertical displacements of upper chord points are larger than the corresponding bottom chord points. Thereby a vertical middle bar put in between the chords promotes an effect of obstruction for normal deformation of contour. A middle bar lightly loaded in compression generates an increase of the bending moment values in the bottom chord sections close to the middle joint significantly (30% up to 35%). The variant of truss with vertical middle bars is analyzed in numerical examples for the illustration the more unfavourable solution and often used in practice.

Table 1: Internal forces and stresses identified in Warren truss elements by different models

<table>
<thead>
<tr>
<th>Structural element</th>
<th>Simplified model</th>
<th>Rigid model (with correct axis of elements at the support)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>axial force, kN</td>
<td>bending moment, kNm</td>
</tr>
<tr>
<td>Support diagonal</td>
<td>586.9 [131.9]</td>
<td>0</td>
</tr>
<tr>
<td>Bottom chord</td>
<td>1367.7 [307.5]</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 1: Structural and design models of the truss: (a) structure, (b) simplified model; (c) rigid model
3 MODELLING OF END-PLATE CONNECTION

3.1 Solution of connection

When the end-plate connection (Fig. 2) has been chosen for tying together the tensioned bottom chord elements made of single I sections, the behaviour of the joint components must be analysed in relationship with tensile force not only but also with their stiffness properties, force distribution and deformations. Due to the coexistence of bending moment and axial force in continuous chord sections there will be regions of distinctive stress concentrations, even though the mean stress across the section remains well below yield.

Here a design model is presented for end-plate connection acceptable for analysis by using common software for engineering calculations (MS Excel, MathCAD or other), and expected results for adequate display of force ranges involved. Thus, the end-plate is modelled as a continuous beam (Fig. 2b) on discrete elastically deformed supports (representing tying bolts), and loaded by concentrated forces transferred by flanges and the web of I section.

![Diagram](image)

Figure 2: Splice joint with extended end-plates (variant I): (a) assembly, (b) loading scheme
3.2 Theoretical background

Solving of end plate problem using the continuous beam theory the assumption that all rotations of beam sections induced by unknown bending moments (internal effort) are equal to ones caused by external loads is applied. As a result of derivations it is obtained that bending moments \( M_{n-1}, M_n, M_{n+1} \) at the three succeeding rigid supports numbered as \((n-1), n, \) and \((n+1)\) of two adjacent spans \( \ell \) of continuous beam are involved by relationship stated by the known Clapeyron’s theorem of three moments [1-3]. This theorem has been used at the first step of the end-plate behaviour analysis assuming bolts as rigid supports. Consequently the first approximation of reactive forces expected to be acting in bolts (as supports) has been determined using the shear force diagram obtained basing on the moment diagram.

In reality some axial deformation of the bolts takes place under tensile forces according to Hooke’s law. Consequently at the next step of analysis a model of a continuous beam on elastic supports (Fig.3) has been examined taking into account the additional rotations \((\psi_{n+1} - \psi_n)\) due to differences of support displacements. The condition of continuity of a deformed axis may be expressed by the following equation:

\[
\frac{\ell}{6EJ} (M_{n-1} + 4M_n + M_{n+1}) + \psi_{n+1} - \psi_n + \tau_{n+1} = 0
\]

where \( \psi_{n+1} \) and \( \psi_n \) are slope angles of beam axis in adjacent spans, that may be expressed through settlements of succeeding supports: \( \psi_n = (u_n-u_{n-1})/\ell, \psi_{n+1} = (u_{n+1}-u_n)/\ell. \)

![Diagram](image)

**Figure 3**: Design model of end-plate (variant I): (a) beam model on discrete elastic supports, (b) proposed deformed axis of end-plate, (c) primary system for analysis
Note that the stiffness of end-plate section \((EI)\) is constant. A new system of three moment equations have to be drawn up and solved in relation to the support moments. Correspondingly, new values of reactions and axial deformations of bolts have been determined.

Analogous calculation cycles may be repeated until bending moment values of two last calculation cycles differ insignificantly. The examples of the procedure are presented below.

### 3.3 Numerical analysis of variant I solution with extended end-plate

An initial design model of end-plate is defined as continuous beam on the rigid supports under loads transferred by flanges and web of bottom chord profiles (see Figs. 3, 4 and Tables 2, 3) assuming elastic material behaviour (see an algorithm in Fig.5).

#### Table 2: Input data for bottom chord joint analysis

<table>
<thead>
<tr>
<th>Name</th>
<th>Symbol, units</th>
<th>Value</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance between flange centres</td>
<td>(h_0), mm [in]</td>
<td>23.75 [0.93]</td>
<td></td>
</tr>
<tr>
<td>Cross section area of profile HEA260</td>
<td>(A_y), mm² [in²]</td>
<td>8680 [13.4]</td>
<td></td>
</tr>
<tr>
<td>Cross section area of flanges</td>
<td>(A_f), mm² [in²]</td>
<td>3496 [5.4]</td>
<td></td>
</tr>
<tr>
<td>Cross section area of web</td>
<td>(A_w), mm² [in²]</td>
<td>1690 [2.6]</td>
<td></td>
</tr>
<tr>
<td>Section modulus about strong axis (W_y)</td>
<td>(W_y), mm³ [in³]</td>
<td>836400 [51]</td>
<td>(W_y = b_p \cdot t_p^2 / 6)</td>
</tr>
<tr>
<td>End-plate section modulus about axis (y_o)</td>
<td>(W_{yo}), mm³ [in³]</td>
<td>26880 [1.14]</td>
<td></td>
</tr>
<tr>
<td>Cross section area of bolt</td>
<td>(A_b), mm² [in²]</td>
<td>314 [0.5]</td>
<td></td>
</tr>
<tr>
<td>Distance between fastener row axis</td>
<td>(\ell), mm [in]</td>
<td>80 [3.15]</td>
<td></td>
</tr>
<tr>
<td>Tensile force in middle bay of bottom chord</td>
<td>(N_{t,d}), kN [in]</td>
<td>1364.1 [306.7]</td>
<td>(\sigma_{t,d} = N_{t,d} / A_y)</td>
</tr>
<tr>
<td>Bending moment</td>
<td>(M_d), kNm [lb*ft]</td>
<td>22.7 [16743]</td>
<td></td>
</tr>
<tr>
<td>Normal stresses caused by axial force</td>
<td>(\sigma_{t,d}), MPa [ksi]</td>
<td>157.2 [22.8]</td>
<td></td>
</tr>
<tr>
<td>Tensile force transferred by upper flange to end-plate</td>
<td>(F_{t,u}), kN [kipf]</td>
<td>454 [102.1]</td>
<td>(F_{t,u} = \sigma_{t,d} A_f \cdot M_d / h_o)</td>
</tr>
<tr>
<td>Tensile force transferred by bottom flange to end-plate</td>
<td>(F_{t,b}), kN [kipf]</td>
<td>645 [145]</td>
<td>(F_{t,b} = \sigma_{t,d} A_f + M_d / h_o)</td>
</tr>
<tr>
<td>Linear tensile load to end-plate</td>
<td>(q_w), N/mm [kipf/ft]</td>
<td>1179 [80.8]</td>
<td>(q_w = \sigma_{t,d} t_w)</td>
</tr>
</tbody>
</table>

![Figure 4: Initial design model of end-plate assuming rigid supports](image)
Figure 5: Algorithm for end-plate joint behavior analysis (variant I)
Table 3: Results of approximation (variant I)

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Support moments, kNm [lbf*ft]</th>
<th>Reactive forces, kN [kipf], (see directions in Fig.4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rigid supports</td>
<td>2nd approximation</td>
</tr>
<tr>
<td>$M_A$</td>
<td>-456.6 [336770]</td>
<td>-384.2 [283371]</td>
</tr>
<tr>
<td>$M_B$</td>
<td>169.8 [125238]</td>
<td>60.5 [44622]</td>
</tr>
<tr>
<td>$M_D$</td>
<td>259.3 [58.3]</td>
<td>273 [61.4]</td>
</tr>
</tbody>
</table>

3.4 Numerical analysis of variant II solution with hidden fasteners

Splice joint solution with hidden end-plates (Fig.6) is an urgent topic under discussion.

Figure 6: End-plate connection of variant II: (a) assembly, (b) loading diagram, (c) primary system for analysis
It is a two span continuous beam model on elastically deformed supports applied for representing of the end-plate behaviour in variant II solution (Fig. 6c). A new three-moment equation (4) has been compiled and solved at the first assumption of the rigid supports, see results presented in Table 4.

\[ \ell \cdot M_B + 4\ell \cdot M_C + \ell \cdot M_D = -6 \cdot R_C^f \]  
\[(2)\]

According to Hooke’s law the values of proposed support settlements have been calculated regarding elastic behaviour in the tension of bolts: \( u_B = 0.0095 \text{ mm}, u_C = 0, u_D = 0.0143 \text{ mm} \). Then a new equation (4) has been drawn up for the beam with elastically settled supports. The results of the calculations are summarized in Table 4. Note that reactive forces \( R_B \) and \( R_D \) are transferred by tensioned bolts, but force \( R_C \) is transferred by end-plate surfaces in local compression.

\[ \left( M_B + 4M_C + M_D \right) \frac{\ell}{6EI} + \frac{u_D - u_C}{\ell} - \frac{u_C - u_B}{\ell} + \tau_C^r + \tau_C^f = 0 \]
\[(3)\]

Equation (4) may be obtained substituting in equation (3) the expression: \( \tau_C^r + \tau_C^f = R_C^f/EI \)

\[ \left( M_B + 4M_C + M_D \right) \frac{\ell}{6EI} + \frac{u_D + u_B}{\ell} + \frac{R_C^f}{EI} = 0 \]
\[(4)\]

**Table 4: Moments and reactive forces for variant II (rigid supports)**

<table>
<thead>
<tr>
<th>Name</th>
<th>Equation</th>
<th>Value for</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Support moments</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( M_B )</td>
<td>(-\frac{F_{tu}}{\ell} \cdot \ell / 2 + q_w \cdot \ell^2 / 8)</td>
<td>-17.0 [12538]</td>
<td>kNm [lbf*ft]</td>
</tr>
<tr>
<td>( M_D )</td>
<td>(-\frac{F_{tb}}{\ell} \cdot \ell / 2 + q_w \cdot \ell^2 / 8)</td>
<td>-28.2 [20799]</td>
<td>kNm [lbf*ft]</td>
</tr>
<tr>
<td>( M_C )</td>
<td>( \frac{M_B \cdot \ell + M_D \cdot \ell + 6 \cdot q_w \cdot \ell^3}{4 \cdot \ell} )</td>
<td>10.5 [7744]</td>
<td>kNm [lbf*ft]</td>
</tr>
<tr>
<td><strong>Support reactions</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( R_B )</td>
<td>(-\frac{1.5F_{tu}}{w} + 1.125q_w\ell \cdot M_C / \ell )</td>
<td>918.5 [206.5]</td>
<td>kN [kipf]</td>
</tr>
<tr>
<td>( R_C )</td>
<td>( \frac{F_{tu} + F_{tb} - 3q_w \ell + F_{tb}}{4\ell} + 2M_C / \ell )</td>
<td>741.7 [166.7]</td>
<td>kN [kipf]</td>
</tr>
<tr>
<td>( R_D )</td>
<td>(-\frac{1.5F_{tb}}{w} + 1.125q_w\ell \cdot M_C / \ell )</td>
<td>1205.1 [270.9]</td>
<td>kN [kipf]</td>
</tr>
<tr>
<td><strong>Elongation of bolts (or settlement of supports)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( u_B )</td>
<td>( \frac{R_B t_p}{[n_f A_{bt} E]} )</td>
<td>0.0095 [0.374]</td>
<td>mm [mil]</td>
</tr>
<tr>
<td>( u_D )</td>
<td>( \frac{R_D t_p}{[n_f A_{bt} E]} )</td>
<td>0.014 [0.551]</td>
<td>mm [mil]</td>
</tr>
</tbody>
</table>
4 SUMMARY

It has been proved by the results of this current study that bending moments generated in the tensile bottom chord sections even of reasonably moderate values itself affects significantly the force distribution between the bolts depending on a location. It has been found that in the case of variant II solution the overloading of bolts near flanges exceed 2,5 times in comparison with the bolts of variant I at the same position, see graphics shown in Fig.7.

No experimental tests have been carried out to prove the results of this study. Significant effects on the behaviour and force distribution in a real structure may be expected due to friction surface-on-surface, imperfections, some lateral actions, plastic deformations possibly accumulated during previous extreme loading cycles and other factors.

Despite the proper account of code conditions in the practical design it would be reasonable to consider the potentials of a structure when subjected to overloading and/or unfavourable service conditions. Progressing the limit state processes may be different in forms and for the intensity of collapse. Evidently, the increase of plastic deformations in compression and tension always takes more time and the rupture in normal bending as well. Only the brittle failure in shear under the thread of the tensioned bolt may happen unexpectedly without any warnings through deformations.

The definition of an adequate structural model is the subject for decision making. Indeed, it is not possible to achieve a complete similarity between the assumed design model and the real structure. However a maximum compliance may be achieved by using modern software in combination with professional judgement because of any value of overload or results from simplified static analysis creates a path to collapse.

![Anticipated force distribution between bolt rows depending on location](image)

Figure 7: Anticipated force distribution between bolt rows depending on location (see positions A, B, C, D, E in Fig.2 and Fig.6)
5 CONCLUSIONS

- Design models that disregard some portion of the stresses or strains in comparison with real structural behaviour, may be assessed as overly simplified and insufficient for the design of bearing structures.
- It is potentially dangerous to assess or design truss structures without thorough analysis of possible models taking into account more unfavourable loading situations, the stresses generated and the deformations developed.
- The continuous beam model on elastically settled supports may be advisable for the analysis of end-plate joint behaviour in heavily loaded tensile chords.
- The end-plate joint solution with fasteners arranged between flanges (variant II) may be assessed as hazardous with regard to collapse.

REFERENCES

Finite Element Analysis of Structural Insulated Panel with OSB Skins against Windborne Debris Impacts

Hong Hao*, Wensu Chen†, Shuyang Chen† and Qingfei Meng†

* Tianjin University and Curtin University Joint Research Center of Structural Monitoring and Protection, School of Civil and Mechanical Engineering, Curtin University
Kent Street, Bentley, WA 6102, Australia
hong.hao@curtin.edu.au

† Tianjin University and Curtin University Joint Research Center of Structural Monitoring and Protection, School of Civil and Mechanical Engineering, Curtin University
Kent Street, Bentley, WA 6102, Australia
wensu.chen@curtin.edu.au

ABSTRACT

Oriented strand board (i.e. OSB) skin structural insulated panel (SIP) with expanded polystyrene (EPS) foam core is becoming more and more widely used as building envelope in the building industry. The structural insulated panel is considered as an efficient panel owing to the advantages of being environmentally sustainable, economical, easy to install and low thermal conductivity, etc. When building envelope is impacted by windborne debris during strong wind event, the debris might perforate the panel, resulting in dominant openings. The dominant openings might cause differential internal pressurization of building and result in building collapse. To investigate the impact resistance capacity, a series of laboratory tests on the SIP panels with OSB skins subjected to a 4kg timber projectile impact were conducted by using a pneumatic cannon system in a previous study. A brief of the experimental tests and results is described first in this paper. This paper then presents a numerical model developed to simulate windborne debris impact on the SIP panels by using commercial software LS-DYNA. The material models of OSB and EPS available in LS-DYNA are incorporated into the numerical model. The accuracy and reliability of the numerical model are validated by comparing the numerical and experimental results in terms of failure modes, penetrated depth of projectile, displacement and strain on the back skin measured in the tests. The validated numerical model can be utilized to conduct more numerical simulations to obtain vulnerability curves and fragility curves of the SIP panel against windborne debris impact.

INTRODUCTION

Structural insulated panel (SIP) is a prefabricated engineered lightweight building material. It consists of a rigid insulating polymer foam core sandwiched by two layers of facial skins. Facial skins can be metal sheet, plywood sheet, Magnesium Oxide board (MgO), oriented strand board (OSB) or fiber cement board etc. [1]. The polymer foam core can be Extruded Polystyrene (XPS) foam, Polyurethane foam or Extended Polystyrene (EPS), etc. In building industry, the structural insulated panels with OSB skin and EPS core have been extensively used for roof, wall and floor sheathing in residential and commercial construction as shown in Figure 1. OSB is a wood composite material manufactured from narrow, long strands bonded together with resin under heat and pressure. It was reported that OSB had been one of the most commonly used wood-based panels for residential construction in North America since 1980s [2].

Strong wind event such as cyclone generated enormous amount of windborne debris and the windborne debris impact was highlighted as a major cause of damage to building envelope [4]. The windborne debris penetration through building envelope typically creates a dominant opening, which results in a differential internal pressurization that contributes to the increased loading of the building’s roof and leeward walls, ultimately may lead to entire roof lifting up or even collapse [5]. Therefore, it is important to understand
the penetration resistance capacity of SIPs, especially when they are used in the regions prone to cyclones as failure is directly related to structural integrity and safety [6]. To withstand the impact of such extreme event, the penetration resistance capacity of wall or roof panels to windborne debris impact should satisfy the testing requirements specified in the respective design codes such as the Australian Wind Loading Code (AS/NZS 1170.2:2011) [7] and FEMA [8]. The performance of various structural panels subjected to projectile impact by using pneumatic cannon system has been studied and numerically simulated in the previous studies [9-13].

In a previous study [10], a series of laboratory tests were carried out by using a pneumatic cannon testing system to investigate the structural response and impact resistance of the structural insulated panel with OSB skin subjected to the timber projectile impact. The failure modes under various impact scenarios were observed. Based on the testing results, a corresponding numerical model is calibrated by employing LS-DYNA in this study. The accuracy of the numerical model is validated by comparing with the testing data in terms of the projectile penetration length, and displacement and strain time histories of the back skin of the specimen.

**BRIEF OF EXPERIMENTAL PROGRAM**

A total of seven specimens i.e. two SIPs with OSB+OSB on both skins and five SIPs with METAL+OSB skins seven SIPs were tested under projectile impacts. The dimension of specimen was 762mm by 1200 mm. The specimen specification and testing scheme are given in Table 1. The skin was made of OSBrace® OSB/3 according to EN300:2006 OSB [14]. The EPS (i.e. extended polystyrene) with class SL and density of 13.5 kg/m³ was utilized as core. The EPS cores with two thicknesses (i.e. 31.6 mm and 43.6 mm) were prepared. Suprasec® 7113 adhesive was used to glue the core and the skins. The testing facilities include pneumatic cannon, 4 kg hardwood projectile with cross-section of dimension 100 mm by 50 mm, two high speed cameras, a chronograph, support frame, laser Linear Voltage Displacement Transducers, strain gauges, data acquisition system, 1500 w halogen lights. The pneumatic actuated cannon launched the wooden projectile at a desired velocity, which was used to simulate the windborne debris. The testing results are also given in Table 1.

<table>
<thead>
<tr>
<th>Specimen</th>
<th>Impact location</th>
<th>Projectile velocity (m/s)</th>
<th>Core thickness (mm)</th>
<th>Thickness of front skin (mm)</th>
<th>Thickness of back skin (mm)</th>
<th>Results</th>
<th>Penetrated length (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1OSB1</td>
<td>center</td>
<td>17</td>
<td>31.6</td>
<td>0.4 (Metal)</td>
<td>6 (OSB)</td>
<td>Penetrated+Stay</td>
<td>68</td>
</tr>
<tr>
<td>1OSB2</td>
<td>center</td>
<td>15</td>
<td>31.6</td>
<td>0.4 (Metal)</td>
<td>6 (OSB)</td>
<td>Penetrated+Stay</td>
<td>12.5</td>
</tr>
<tr>
<td>1OSB3</td>
<td>center</td>
<td>17</td>
<td>31.6</td>
<td>6 (OSB)</td>
<td>0.4 (Metal)</td>
<td>Penetrated+Stay</td>
<td>14</td>
</tr>
<tr>
<td>1OSB4</td>
<td>center</td>
<td>17</td>
<td>43.6</td>
<td>0.4 (Metal)</td>
<td>6 (OSB)</td>
<td>Passed+cracked</td>
<td>-</td>
</tr>
<tr>
<td>1OSB5</td>
<td>quarter</td>
<td>17</td>
<td>43.6</td>
<td>0.4 (Metal)</td>
<td>6 (OSB)</td>
<td>Penetrated+Stay</td>
<td>24</td>
</tr>
</tbody>
</table>
NUMERICAL SIMULATIONS

The specimen named 2OSB1 with EPS core thickness 31.6 mm and both OSB skins 6 mm is selected to calibrate the numerical model. A numerical model is developed to simulate windborne debris impact on sandwich panel in LS-DYNA. The failure modes and the structural responses including projectile penetrated length, displacement and strain are used to examine the accuracy of the numerical model.

Finite element model

The finite element model of the specimen is depicted in Figure 2. Since the panel is symmetric about X-Y and Y-Z planes, only a quarter of the panel was modeled to reduce the model size. The model consists of 92,241 solid elements. Both OSB skin and EPS core are meshed by using eight-node solid element with single integration points (i.e. ELFORM=1, constant stress solid elements) to overcome the negative volume, which is prone to happen in full integration elements. As the numerical simulation result is sensitive to mesh size, convergence study was carried out by using three mesh sizes of 10 mm, 5 mm and 2.5 mm in the vicinity of impact where dense elements are applied. The numerical results show the predicted projectile penetration depth by using mesh size of 2.5 mm yields close results to the actual projectile depth recorded in the test. Therefore, the solid element with element size of 2.5 mm is used for both OSB skins and EPS foam core in the region of impact where dense elements are applied. Other region uses 15 mm mesh size where non-dense elements are applied.

In the test, the specimen was clamped on the support frame by using G-clamps along four sides as shown in Figure 2. In the numerical model, the specimen is assumed constrained on three DOFs of UX, UY and UZ at two boundary edges by using *BOUNDARY SPC SET. The other two sides are symmetrical about Y-Z and X-Y planes, respectively.

Contact algorithm

The contact between the projectile and the specimen is defined by using *CONTACT ERODING SURFACE TO SURFACE with segment based contact option (i.e. SOFT=2). Scale factors on penalty stiffness are increased to avoid the excessive penetration of the slave part into master part. To account for the contact between OSB skins and core, *CONTACT ERODING SINGLE SURFACE is used and *CONTACT INTERIOR is utilized to avoid the negative volume within the EPS core. De-bonding
between the skins and the core is not obvious in the tests. Therefore, the contact between OSB skin and EPS foam is assumed perfectly bonded and modelled as common points. The coefficients of static and dynamic friction between the projectile and the materials are assumed to be 0.5 and 0.4, respectively.

Material model

The 4kg timber projectile is modeled as a rigid body as the hardwood projectile experienced no obvious deformation and mass loss after the tests. As EPS exhibits strain rate dependency, the material model *MAT MODIFIED CRUSHABLE FOAM (MAT_163), which incorporates the strain rate effect into the material model *MAT CRUSHABLE FOAM (MAT_63), is used for EPS core. To avoid the negative volume, hourglass control with assumed strain co-rotational stiffness form (i.e. IHQ=6) is used. The material properties of EPS with density of 13.5 kg/m³ are given as follows: E=890 kPa, Tc=40kPa, Poisson’s ratio=0, maximum shear strain (MSS) at failure=0.04, which were obtained in laboratory tests [15] and calibrated in the previous study [9].

A transversely isotropic material model *MAT WOOD (MAT_143) is adopted for the modelling of OSB/3 [16, 17]. The basic properties of OSB/3 such as density, elastic modulus, flexure strength are provided by the supplier. Other material properties such as stiffness and strength are referred to the standards and literature as given and compared in Table 2. The value in the bracket stands for the parameter in the perpendicular axis. Due to the complexity nature of OSB material, the numerical model was calibrated by means of extensive trial and error. The material parameters of *MAT WOOD with respect to stiffness, strength and damage [17] are determined in Table 3. The remaining variables in the Keyword use default values.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Density (kg/m³)</td>
<td>638</td>
<td>550</td>
<td>638</td>
<td>0.16–0.23</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>Poisson’s ratio</td>
<td>≥4.5 (1.8)</td>
<td>3.5 (1.4)</td>
<td>4.8–8.3</td>
<td>4.5 (1.8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elastic Modulus</td>
<td>1.08</td>
<td>1.2–2</td>
<td>1.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shear Modulus thru thickness</td>
<td>0.05</td>
<td>0.14–0.34</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shear Modulus in-plane</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flexure strength</td>
<td>≥22 (11)</td>
<td>22 (11)</td>
<td>18 (9)</td>
<td>20.7–27.6</td>
<td>22(11)</td>
<td></td>
</tr>
<tr>
<td>Compressive Strength</td>
<td>15.9 (12.9)</td>
<td>10.3–17.2</td>
<td>15.9 (12.9)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shear Strength thru thickness</td>
<td>6.8</td>
<td>6.9–10.3</td>
<td>6.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shear Strength in-plane</td>
<td>1.0</td>
<td>1.4–2.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Material properties of OS'Brace ® OSB/3 (from supplier, standards and literature)

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Variable</th>
<th>Value</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stiffness:</td>
<td>EL</td>
<td>4.5</td>
<td>GPa</td>
</tr>
<tr>
<td>Parallel normal modulus</td>
<td>ET</td>
<td>1.8</td>
<td>GPa</td>
</tr>
<tr>
<td>Perpendicular normal modulus</td>
<td>GLT</td>
<td>1.1</td>
<td>GPa</td>
</tr>
<tr>
<td>Perpendicular shear modulus</td>
<td>GTR</td>
<td>1.1</td>
<td>GPa</td>
</tr>
<tr>
<td>Parallel shear modulus</td>
<td>vLT</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>Strength:</td>
<td>XT</td>
<td>9.9</td>
<td>MPa</td>
</tr>
<tr>
<td>Property</td>
<td>Symbol</td>
<td>Value 1</td>
<td>Value 2</td>
</tr>
<tr>
<td>--------------------------------------</td>
<td>--------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>Parallel compressive strength</td>
<td>XC</td>
<td>15.9</td>
<td>MPa</td>
</tr>
<tr>
<td>Perpendicular tensile strength</td>
<td>YT</td>
<td>7.2</td>
<td>MPa</td>
</tr>
<tr>
<td>Perpendicular compressive strength</td>
<td>YC</td>
<td>12.9</td>
<td>MPa</td>
</tr>
<tr>
<td>Parallel shear strength</td>
<td>S</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perpendicular shear strength</td>
<td>S⊥</td>
<td>6.8</td>
<td>MPa</td>
</tr>
</tbody>
</table>

**Damage:**

<table>
<thead>
<tr>
<th>Property</th>
<th>Symbol</th>
<th>Value 1</th>
<th>Value 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parallel fracture energy in tension</td>
<td>Gf I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parallel fracture energy in shear</td>
<td>Gf II</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parallel softening parameter</td>
<td>B</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>Parallel maximum damage</td>
<td>dmax</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perpendicular fracture energy in tension</td>
<td>Gf I ⊥</td>
<td>6.2</td>
<td>kJ/m²</td>
</tr>
<tr>
<td>Perpendicular fracture energy in shear</td>
<td>Gf II ⊥</td>
<td>6.2</td>
<td>kJ/m²</td>
</tr>
<tr>
<td>Perpendicular softening parameter</td>
<td>D</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>Perpendicular maximum damage</td>
<td>dmax⊥</td>
<td>0.99</td>
<td></td>
</tr>
</tbody>
</table>

**COMPARISONS OF EXPERIMENTAL AND NUMERICAL RESULTS**

The failure modes and the structural responses including penetrated length, displacement and strain are used to examine the accuracy of numerical simulation. As observed from the testing, the specimen 2OSB1 was penetrated with the penetrated length of 26 cm by the projectile at the velocities of 10 m/s. The specimen 2OSB1 experienced skin shear punch and core shear failure with a clean rectangular hole. The failure mode in the numerical analysis is comparable to that obtained in the test as shown in Figure 3 and Figure 4. The penetrated length retrieved from the numerical results is 22 cm which is closed to 26 cm measured in the test. Figure 5 shows the Von Mises stress distributions. It is found that high stress occurs near the vicinity of impact. The LVDT measures time histories at the center point on the back skin. The numerical simulated displacement time history captures the trend of the recorded displacement response as shown Figure 6. The peak displacements obtained from numerical simulation and experimental test are 9.4 mm and 9.0 mm, respectively, with an error of 4.4%. As shown in Figure 6, the numerically simulated strain time history of the point at the mid-way between the center and the edge of the panel also captures the trend of that recorded in experimental test. Through comparison of numerical results and testing results, it can be concluded that the numerical model gives reasonably accurate predictions of the panel responses to the projectile impact.

**Figure 3:** (L) Numerical result; (R) Experimental result (Front View-2OSB1)
Hong Hao, Wensu Chen, Shuyang Chen and Qingfei Meng

Figure 4: (L) Numerical result (Reflected to full model); (R) Experimental result (Back View-2OSB1)

Figure 5: Von Mises stress contours of specimen (Reflected to full model)

Figure 6: (L) Comparison of displacement-time history from numerical and test results; (R) Comparison of strain-time history from numerical and test results

CONCLUSION

This paper presents a numerical model developed to simulate the response of structural insulated panel with OSB skins subjected to projectile impact by using LS-DYNA. The numerical model has been calibrated with the test results. The validated numerical model can be used to conduct intensive simulations to generate vulnerability curves and fragility curves of the structural insulated panel under windborne debris impact, which may be useful for the probabilistic loss estimations of structural panels and building
structures in strong wind events.

ACKNOWLEDGMENTS
The authors acknowledge the financial support from Australian Commonwealth Scientific and Industrial Research Organization (CSIRO) through the project “Climate Adaptation Engineering for Extreme Events Cluster”.

REFERENCES
ULTIMATE LOADING CAPACITY OF RC COLUMNS UNDER FIRE: COMPARISON OF ACI AND EUROCODE MODELS

H. BARROS¹, C. FERREIRA¹ AND V. D. SILVA¹

¹ Institute for Systems Engineering and Computers at Coimbra (INESC Coimbra)
Dep. of Civil Engineering, Faculty of Sciences and Technology, University of Coimbra, Portugal
hbarros;carla;vdsilva@dec.uc.pt

Key words: Reinforced Concrete, Columns, Thermal Properties, Eurocode 2, ACI, Fire Analysis.

Abstract. The reinforced concrete structures undergo a loss of capacity when submitted to the high temperatures due to a fire. The mechanical properties namely the strength of concrete and steel are predicted in different ways in the European codes (EC2) [1] and in the American Concrete Institute (ACI) code [2].
In the present work the temperature after 30 minutes and 60 minutes in columns under standard fire exposure ISO 834 are obtained. The software implemented considers geometric and material nonlinearities, obtaining the load displacement curves to compare the different models of EC2 and ACI.

1 INTRODUCTION

The reinforced concrete structures undergo a loss of capacity when submitted to high temperatures because the mechanical properties, namely the strength of concrete and steel, are substantially affected by the temperature raise. The computation of the load capacity needs the evaluations of temperature within the structure and the change in properties. Laboratorial tests conducted in concrete specimens under high temperatures and different loading conditions have led the European codes, EC2 [1], and in the American Concrete Institute code, ACI [2], to different models of stress strain temperature dependent relations. The EC2 [1] recommends temperature dependent curves distinct for siliceous and calcareous aggregates. The ACI code [2], besides this distinction takes into account the loading condition considering three cases: 1-specimens stressed during the heating; 2-unstressed during the heating; 3-stressed after attaining ambient temperature, termed residual.
In this work the ultimate load capacity of columns under fire is computed using ACI and EC2 models and the values are compared. Temperature due to fire is obtained in a thermal transient analysis solved by the finite element method. The columns with rectangular section, siliceous aggregates, are exposed to standard fire exposure ISO 834 during 30 minutes and 60 minutes. The thermal analysis considers the lower limit of the conductivity indicated in EC2 for this kind of aggregates.
The ultimate load capacity of the columns was computed by a numerical method developed in [4]. Both material and geometrical non-linearity are considered.
displacement method with a classical Newton-Raphson algorithm is used for the computation of the equilibrium configurations. Full interaction between internal forces and deformation is taken into account, in order to fully enforce the equilibrium conditions in the deformed configuration. Arbitrary large rotations and displacements are considered. The deformations are, however, considered as small. The description of the geometry is of the totally Lagrangian type. A bar finite element is used.

2 MATERIAL PROPERTIES AT ELEVATED TEMPERATURES

The material properties of steel and concrete, namely the parameters of the stress-strain law, are defined according to the local temperature. This means that at each fibre of the section, with a known temperature, the stress-strain law is different from a nearby fibre with different temperature. In the next sections the relevant properties of steel and concrete are described.

2.1 Steel

As far as the reinforcing steel is concerned it has a nonlinear constitutive relation (see Eurocode 2 [1]) for tension and compression reinforcement, function of temperature \( \theta \), defined by the following parametric equation (disregarding the descending branch):

\[
\sigma = \begin{cases} 
\varepsilon E_{s,\theta} & \text{if } \varepsilon \leq \varepsilon_{s,\theta} \\
 f_{sp,\theta} - \varepsilon + \left(\frac{b}{a}\right) \left( a^2 - \left(\varepsilon_{sy,\theta}\right)^2 \right)^{0.5} & \text{if } \varepsilon_{s,\theta} \leq \varepsilon \leq \varepsilon_{sy,\theta} \\
 f_{sy,\theta} & \text{if } \varepsilon_{sy,\theta} \leq \varepsilon \leq \varepsilon_{st,\theta}
\end{cases}
\]

The parameters in this equation are found in Eurocode 2 [1], that are: \( f_{yk} \) yield strength; \( f_{sy,\theta} \) yield strength for temperature equal to \( \theta \); \( \varepsilon_{s,\theta} \); \( \varepsilon_{sy,\theta} \); \( \varepsilon_{st,\theta} \); \( \varepsilon_{su,\theta} \) strains, at temperature \( \theta \).

2.2 Normal concrete EC2

The concrete stress \( \sigma \) is defined in terms of the strain \( \varepsilon \) in Eurocode 2 [1] by a parametric function that is:

\[
\sigma = \begin{cases} 
\frac{3 \varepsilon f_{c,\theta}}{\varepsilon_{c1,\theta} \left( 2 + \left( \frac{\varepsilon}{\varepsilon_{c1,\theta}} \right)^3 \right)} & \text{if } \varepsilon \leq \varepsilon_{c1,\theta} \\
\text{descending branch if } \varepsilon_{c1,\theta} \leq \varepsilon \leq \varepsilon_{cu1,\theta}
\end{cases}
\]

The parameters in this equation are: \( f_{ck} \) compressive strength at ambient temperature; \( f_{c,\theta} \) compressive strength at temperature \( \theta \); \( \varepsilon_{c1,\theta} \) strain at peak stress at temperature \( \theta \); \( \varepsilon_{cu1,\theta} \) ultimate strain at temperature \( \theta \). The parameters needed are resumed in table 1, following Eurocode 2 [1] recommendations.
Table 1: Values for the main parameters of the stress-strain relationships of normal concrete.

<table>
<thead>
<tr>
<th>Concrete temperature $\theta$ [°C]</th>
<th>Siliceous aggregate</th>
<th>Calcareous aggregate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$f_{c,0}/f_{ck}$</td>
<td>$\varepsilon_{c,0}$</td>
</tr>
<tr>
<td>20</td>
<td>1.00</td>
<td>0.0025</td>
</tr>
<tr>
<td>100</td>
<td>1.00</td>
<td>0.0040</td>
</tr>
<tr>
<td>200</td>
<td>0.95</td>
<td>0.0055</td>
</tr>
<tr>
<td>300</td>
<td>0.85</td>
<td>0.0070</td>
</tr>
<tr>
<td>400</td>
<td>0.75</td>
<td>0.0100</td>
</tr>
<tr>
<td>500</td>
<td>0.60</td>
<td>0.0150</td>
</tr>
<tr>
<td>600</td>
<td>0.45</td>
<td>0.0250</td>
</tr>
<tr>
<td>700</td>
<td>0.30</td>
<td>0.0250</td>
</tr>
<tr>
<td>800</td>
<td>0.15</td>
<td>0.0250</td>
</tr>
<tr>
<td>900</td>
<td>0.08</td>
<td>0.0250</td>
</tr>
<tr>
<td>1000</td>
<td>0.04</td>
<td>0.0250</td>
</tr>
<tr>
<td>1100</td>
<td>0.01</td>
<td>0.0250</td>
</tr>
<tr>
<td>1200</td>
<td>0.00</td>
<td>–</td>
</tr>
</tbody>
</table>

2.3 Comparison of compressive strength concrete models of ACI and EC2

Different concrete stress-strain laws in compression are proposed by ASCE [5], Popovics [7], Kodur et al. [7], and Eurocode [1]. In ACI the compressive thermal-dependent models of concrete are based in research involved 3-in x 6-in (7.6-cm x 15.2-cm) cylindrical specimens that were subjected to three test scenarios: stressed during heating, unstressed during heating and residual tests. In last section the properties of unstressed scenario are considered and are used in the numerical applications. Different from the ACI 216 guidelines, the Eurocode and CEB models do not consider the test type as a parameter. Fig. 1 presents the temperature-dependent compressive strength models for EC 2 and the three scenarios of ACI.

![Figure 1: EC 2 and ACI temperature-dependent compressive strength models.](image-url)
3 NUMERICAL EXAMPLES

3.1 Temperature distribution in the cross section

The concrete used in the columns is made of siliceous aggregates. The thermal analysis considers the lower limit of the conductivity indicated in EC2 for this kind of aggregates. The temperature distribution is indicated in Fig 2a). The cross section geometry and column schema are represented in Fig 2b) and c), respectively. Three different amount of reinforcement $A_s$ are considered in the examples: 28.14cm$^2$ (2 x 7φ16); 43.98cm$^2$ (2 x 7φ20); 78.54cm$^2$ (2 x 8φ25).

![Temperature distribution in the cross section after 30 minutes of exposure to fire in all sides](image1)

**Figure 2:** a) Temperature distribution in the cross section after 30 minutes of exposure to fire in all sides (a quarter of the section) b) cross section and c) column geometries.

3.2 Behaviour of the columns

The ultimate loads are found for three lengths of column (3.0m, 6.0m and 9.0m) with two eccentricities (18cm and 27cm) considered in major inertia axis (see Fig. 2c)), at the end of 30 min and 60 min of fire. The tables 2 to 4 resume the results obtained with the ACI and EC2 codes for the three reinforcement areas.
4 CONCLUSIONS

- In all columns and temperature configurations the ultimate load obtained with ACI is smaller than with EC2.
- The loss of capacity increases when the length of the column decreases and this effect is more relevant in ACI than EC2.
- The difference in the loss of capacities between the two codes is not modified by the increase of the first order bending moment.

Table 2: Ultimate load (kN) in columns with 2 x 7\(\phi\)16 reinforcement.

<table>
<thead>
<tr>
<th></th>
<th>L = 3 m</th>
<th>L = 6 m</th>
<th>L = 9 m</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>e = 18 cm</td>
<td>e = 27 cm</td>
<td>e = 9 cm</td>
</tr>
<tr>
<td>ACI</td>
<td>5715</td>
<td>4329</td>
<td>6566</td>
</tr>
<tr>
<td>EC2</td>
<td>5715</td>
<td>4332</td>
<td>6566</td>
</tr>
<tr>
<td>(\theta=20^\circ\text{C})</td>
<td>30&quot; fire ((f_{30}))</td>
<td>4770</td>
<td>5051</td>
</tr>
<tr>
<td></td>
<td>60&quot; fire ((f_{60}))</td>
<td>4141</td>
<td>4483</td>
</tr>
<tr>
<td>Loss of capacity</td>
<td>30&quot; fire ((f_{30}))</td>
<td>4495</td>
<td>4680</td>
</tr>
<tr>
<td></td>
<td>60&quot; fire ((f_{60}))</td>
<td>3750</td>
<td>4100</td>
</tr>
</tbody>
</table>

Table 3: Ultimate load (kN) in columns with section 2 x 7\(\phi\)20 reinforcement.

<table>
<thead>
<tr>
<th></th>
<th>L = 3 m</th>
<th>L = 6 m</th>
<th>L = 9 m</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>e = 18 cm</td>
<td>e = 27 cm</td>
<td>e = 9 cm</td>
</tr>
<tr>
<td>ACI</td>
<td>6325</td>
<td>4937</td>
<td>7194</td>
</tr>
<tr>
<td>EC2</td>
<td>6325</td>
<td>4937</td>
<td>7194</td>
</tr>
<tr>
<td>(\theta=20^\circ\text{C})</td>
<td>30&quot; fire ((f_{30}))</td>
<td>5390</td>
<td>4327</td>
</tr>
<tr>
<td></td>
<td>60&quot; fire ((f_{60}))</td>
<td>4724</td>
<td>3830</td>
</tr>
<tr>
<td>Loss of capacity</td>
<td>30&quot; fire ((f_{30}))</td>
<td>4496</td>
<td>3580</td>
</tr>
<tr>
<td></td>
<td>60&quot; fire ((f_{60}))</td>
<td>3750</td>
<td>2809</td>
</tr>
</tbody>
</table>

Table 4: Ultimate load (kN) in columns with section 2 x 8\(\phi\)25 reinforcement.

<table>
<thead>
<tr>
<th></th>
<th>L = 3 m</th>
<th>L = 6 m</th>
<th>L = 9 m</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>e = 18 cm</td>
<td>e = 27 cm</td>
<td>e = 9 cm</td>
</tr>
<tr>
<td>ACI</td>
<td>7550</td>
<td>6060</td>
<td>8517</td>
</tr>
<tr>
<td>EC2</td>
<td>7550</td>
<td>6060</td>
<td>8517</td>
</tr>
<tr>
<td>(\theta=20^\circ\text{C})</td>
<td>30&quot; fire ((f_{30}))</td>
<td>6594</td>
<td>5394</td>
</tr>
<tr>
<td></td>
<td>60&quot; fire ((f_{60}))</td>
<td>5825</td>
<td>4809</td>
</tr>
<tr>
<td>Loss of capacity</td>
<td>30&quot; fire ((f_{30}))</td>
<td>5427</td>
<td>4244</td>
</tr>
<tr>
<td></td>
<td>60&quot; fire ((f_{60}))</td>
<td>4809</td>
<td>3629</td>
</tr>
</tbody>
</table>
The previous conclusion is also observed for the increase of reinforcement area.

ACKNOWLEDGEMENTS

The financial support of ACIV is gratefully acknowledged.

REFERENCES

VIBRATION ANALYSIS AND HUMAN COMFORT ASSESSMENT OF STEEL-CONCRETE COMPOSITE FLOORS SUBJECTED TO HUMAN RHYTHMIC ACTIVITIES

CÁSSIO M. R. GASPAR¹, JOSÉ GUILHERME S. DA SILVA²

¹Civil Engineering Post-graduate Programme, PGECIV, State University of Rio de Janeiro, UERJ
São Francisco Xavier St., 524, Maracanã, Rio de Janeiro/RJ, Brazil, CEP: 20550-900
E-mail: cgaspac@yahoo.com.br

²Structural Engineering Department, ESTR, State University of Rio de Janeiro, UERJ
São Francisco Xavier St., 524, Maracanã, Rio de Janeiro/RJ, Brazil, CEP: 20550-900
E-mail: jgss@uerj.br

Key words: Human Comfort, Excessive Vibrations, Dynamic Analysis, Composite Floors.

Abstract. Steel-concrete composite floor vibrations induced by human rhythmic activities like walking, running, jumping or aerobics consist on a very complex problem. The dynamic action characteristics generated during these human activities are directly related to the individual body and to the specific way in which each person executes a certain rhythmic activity. The main objective of this research is to investigate the dynamic behaviour of a steel-concrete composite floor subjected to rhythmic human activities (aerobics). The investigated structural model is based on a steel-concrete composite floor spanning 40m by 40m, with a total area of 1600m² and represents a typical interior floor bay of a commercial building used for gym. In this research, an extensive parametric study was developed aiming to obtain the peak accelerations, RMS and VDV values, based on five different mathematical formulations used for modelling human rhythmic actions (aerobics). The dynamic response of the investigated structural model have indicated that the steel-concrete composite floor presented high vibration levels that compromise the human comfort, according to the limiting values proposed by several authors and design standards.

1 INTRODUCTION

Structural engineers are constantly faced with an economical design versus the structural safety and serviceability in a general design process. In fact, considering the design of steel-concrete composite floors under human rhythmic activities, one may use some force functions in order to represent the human activity acting on the floor. Namely, some design standards, practical design guides and authors as ISO10137 [1], CEB [2, 3], SCI [4], AISC [5] and FAISCA [6] might be considered.

On the other hand, taking into account that the mathematical representation of human rhythmic activities consists on a very complex problem, most of these force functions does not agree to each other. Consequently, the choice of one or other force function may lead the floor design to a more or less conservative assessment regarding the human comfort serviceability. Besides, it is important to highlight that beyond the design stage evaluation of
composite floors, the choice of one of these force functions is likewise significant and plays an important role in relation to already built floors assessment.

In parallel, it must be emphasized that several authors have been related that the current composite floors design might be susceptible to the resonance phenomenon, causing undesirable vibrations in the frequency range that is the most noticeable to humans, i.e. 4 to 8Hz [3, 4, 7]. Such condition is in part due to the technological advance in the materials field which has allowed the use of more resistant and low weight materials that result in slender and more flexible structural floor systems. In addition, it has been observed in design practice low floor structural damping ratios, which is related to the type of construction, materials, presence of non-structural elements, age and quality of construction [1]. Therefore, all these combinations make the steel-concrete composite floors very susceptible to annoying vibrations induced by humans.

This way, the main objective of this research is to investigate the dynamic behaviour of a steel-concrete composite floor subjected to rhythmic human activities (aerobics). The investigated structural model is based on a steel-concrete composite floor spanning 40m by 40m, with a total area of 1600m² and represents a typical interior floor bay of a commercial building used for gym.

In this research, an extensive parametric study was developed aiming to obtain the peak accelerations, RMS and VDV values, based on five different mathematical formulations used for modelling human rhythmic actions (aerobics). The dynamic response of the investigated structural model have indicated that the steel-concrete composite floor presented high vibration levels that compromise the human comfort, according to the limiting values proposed by several authors and design standards.

2 FINITE ELEMENT MODELLING

The finite element model was based on a real steel-concrete composite floor spanning 40m by 40m with a total area of 1600m² composed by 16 bays of 10x10m. The structural system is made from composite beams and a 100mm thick concrete slab, see Figure 1. The columns height is equal to 4m. The steel sections used were welded wide flanges (WWF) made from a 345MPa yield stress steel grade. A 205x10³MPa Young’s modulus was adopted for the steel beams. The concrete slab has a 30MPa specified compression strength and a 26x10³ MPa Young’s Modulus.

The proposed computational model adopted the usual mesh refinement techniques present in finite element method simulations implemented in the ANSYS program [8], see Figure 2. The floor steel beams were represented by three-dimensional beam elements, where flexural and torsion effects are considered. The composite slab was represented by shell finite elements. Both materials (steel and concrete) have an elastic behaviour.

The complete interaction between the concrete slab and steel beams was considered in the analysis and simulated by spring elements so that the numerical model coupled all the nodes between the beams and slab to prevent the occurrence of any slip. Regarding the behaviour of the structural connections present in the investigated composite floor, the beam-to-beam connections were modelled as flexible joints using spring elements. The beam-to-column connections were considered as rigid joints.
The final numerical-computational model was made out of 29874 nodes, 3920 three-dimensional beam elements and 25600 shell elements, resulting in a numerical model with 167069 degrees of freedom.

3 RHYTHMIC ACTIVITIES MODELLING

The representative mathematical function of the rhythmic dynamic loading can be described by five different experimental approaches such as proposed by ISO10137 [1], CEB [2, 3], SCI [4], AISC [5] and FAISCA [6]. The Figures 3 and 4 show an example of the dynamic force in the time and frequency domain considering a frequency step of 2.57Hz. It is also assumed that a single person’s weight is 800N.
Figure 3: Dynamic loading functions in time domain

a) The first three harmonics of the force functions

b) Third harmonic of the force functions

Figure 4: Dynamic loading functions in frequency domain

Table 1: Force frequency range corresponding to the third harmonic of aerobic activities

<table>
<thead>
<tr>
<th>Dynamic loading model</th>
<th>Frequency range related to the third harmonic of aerobics (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAISCA [6]</td>
<td>5.66 - 8.57</td>
</tr>
<tr>
<td>SCI [4]</td>
<td>4.5 - 8.4</td>
</tr>
<tr>
<td>CEB [2, 3]</td>
<td>5.4 - 10.2</td>
</tr>
<tr>
<td>ISO 10137 [1]</td>
<td>4.5 - 10.5</td>
</tr>
</tbody>
</table>
From Figure 4 one can see that the third Fourier coefficients related to the CEB and ISO 10137 functions are more energetic than the SCI, AISC and FAISCA functions ones. Therefore, this fact will be investigated in this work. Table 1 presents the frequency range related to the third harmonic of aerobics activities. It can be noted that the CEB and ISO 10137 are more conservative in relation to the upper limit frequency range.

4 STEEL-CONCRETE COMPOSITE FLOOR DYNAMIC ANALYSIS

4.1 Natural Frequencies and Vibration Modes

The modal analysis was carried out and it was verified that the first twenty-four natural frequencies of the analysed floor, varying from 6.06Hz to 10.32Hz, are close to the excitation frequency range (aerobics). In this situation, the frequency of the third harmonic of the dynamic loading, varying according to Table 1, may match these natural frequencies and therefore lead the composite floor to resonance. Consequently, such situation might result in undesirable vibrations and thus human annoyance.

<table>
<thead>
<tr>
<th>Frequencies $f_{0i}$(Hz)</th>
<th>Composite Floor Frequencies</th>
<th>Frequencies $f_{0i}$(Hz)</th>
<th>Composite Floor Frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_{01}$</td>
<td>6.06</td>
<td>$f_{14}$</td>
<td>7.62</td>
</tr>
<tr>
<td>$f_{02}$</td>
<td>6.36</td>
<td>$f_{15}$</td>
<td>7.72</td>
</tr>
<tr>
<td>$f_{03}$</td>
<td>6.44</td>
<td>$f_{16}$</td>
<td>7.79</td>
</tr>
<tr>
<td>$f_{04}$</td>
<td>6.65</td>
<td>$f_{17}$</td>
<td>8.90</td>
</tr>
<tr>
<td>$f_{05}$</td>
<td>6.92</td>
<td>$f_{18}$</td>
<td>9.30</td>
</tr>
<tr>
<td>$f_{06}$</td>
<td>7.05</td>
<td>$f_{19}$</td>
<td>9.38</td>
</tr>
<tr>
<td>$f_{07}$</td>
<td>7.11</td>
<td>$f_{20}$</td>
<td>9.63</td>
</tr>
<tr>
<td>$f_{08}$</td>
<td>7.12</td>
<td>$f_{21}$</td>
<td>10.09</td>
</tr>
<tr>
<td>$f_{09}$</td>
<td>7.13</td>
<td>$f_{22}$</td>
<td>10.12</td>
</tr>
<tr>
<td>$f_{10}$</td>
<td>7.27</td>
<td>$f_{23}$</td>
<td>10.23</td>
</tr>
<tr>
<td>$f_{11}$</td>
<td>7.35</td>
<td>$f_{24}$</td>
<td>10.32</td>
</tr>
<tr>
<td>$f_{12}$</td>
<td>7.44</td>
<td>$f_{25}$</td>
<td>10.77</td>
</tr>
<tr>
<td>$f_{13}$</td>
<td>7.55</td>
<td>$f_{26}$</td>
<td>10.83</td>
</tr>
</tbody>
</table>
Figure 5: Investigated structural model vibration modes

Figure 5 shows some vibration modes which might be excited by the dynamic load positioning. It is must be noted that all floor vibration modes present a predominant flexural behaviour. With the objective of avoiding unwanted vibrations the fundamental floor frequency should be at least equal to 9.2 Hz (AISC [5]) for aerobics considering the structural system as a light floor or 8.4 Hz for rhythmic activities according to SCI [4].

4.2 Human Comfort Assessment

The dynamic transient analysis was performed considering sixty-four individuals practising aerobics (0.25 person/m² [3]) so that the third harmonic of each force function matches the fifteenth vibration mode \( f_{15} = 7.72 \text{ Hz} \) aiming to cause a resonant motion state (see Figure 6). It is also assumed that a single person’s weight is 800N (0.8kN).

The modal damping ratio of 1% was chosen according to ISO 10137 [1] and lies in the extreme range of 0.5% to 8.0% for fully composite steel beams with shear connectors to concrete slab. The dynamic response of the composite floor was obtained from the central node of each floor bay where the rhythmic activity is taken place, which is represented by the letters F, G, and K (see Figure 6). The steel-concrete composite floor accelerations herein were evaluated based on peak unweighted accelerations (steady-state response), \( a_{p} \), RMS weighted accelerations, \( a_{w,rms} \), and vibration dose values, VDV. Table 3 shows the unweighted acceleration, RMS weighted accelerations and the VDV acceptance criteria respectively.
Cássio M. R. Gaspar, José Guilherme S. da Silva

a) Sixty-four individuals practising aerobics on the floor

b) Investigated sections of the composite floor

Figure 6: Persons’ positioning along the floor. Dimensions in (m)

Table 3: Human comfort acceptance criteria

<table>
<thead>
<tr>
<th>Dynamic Loading Model</th>
<th>Person’s Reaction</th>
<th>a_p (m/s²)</th>
<th>a_w,rms (m/s²)</th>
<th>VDV (m/s^{1.75})</th>
<th>Person’s Reaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>AISC [5]</td>
<td>Reasonable limit: passive persons</td>
<td>&lt; 0.35</td>
<td>&lt; 0.66</td>
<td>&lt; 0.50</td>
<td></td>
</tr>
<tr>
<td>SCI [4]</td>
<td>Disturbing</td>
<td>0.35 - 1.27</td>
<td>0.66 - 2.38</td>
<td>0.50 - 3.50</td>
<td></td>
</tr>
<tr>
<td>Ellis and Littler [9]</td>
<td>Unacceptable/ Panic</td>
<td>1.27 - 2.47</td>
<td>2.38 - 4.64</td>
<td>3.50 - 6.90</td>
<td></td>
</tr>
<tr>
<td>SCI [4]</td>
<td>Unacceptable/ Panic</td>
<td>&gt; 2.47</td>
<td>&gt; 4.64</td>
<td>&gt; 6.90</td>
<td>Probably causing panic</td>
</tr>
</tbody>
</table>

From a comparison between Tables 3 and 4, it can be observed that all acceleration values are greater than the human comfort criteria reasonable limits (Table 3) for all cases. Indeed, such condition might be expected once the natural frequencies and vibration modes of the composite floor lie in the third harmonic of the force functions ranges. Therefore, the sixty-four individuals who were practising aerobics might be subjected to extremely annoying vibrations. Figures 7 and 8 present the steel-concrete composite floor dynamic response in time domain and frequency domain.
Figure 7: Investigated steel-concrete composite floor time domain response

Figure 8: Investigated steel-concrete composite floor frequency response
With respect to Figure 7, one can note the steady-state acceleration in the time domain from 5s. In relation to Figure 8, it can be seen that the ISO 10137 [1] and CEB [2] force functions presented the third harmonic more energetic than those compared to SCI, AISC and FAISCA [6]. Note that this result could be expected from Figure 4.

In general, the accelerations obtained by the ISO 10137 [1] and CEB [2] functions were very conservative in terms of the Fourier coefficients corresponding to the third harmonic. In fact, CEB [2] was the most conservative one. The SCI [4] and AISC [5] functions presented accelerations at the same level whereas the lowest accelerations values were obtained according to the FAISCA [6] function. However, the acceleration results from FAISCA [6] were still found in disturbing vibration levels.

5 CONCLUSIONS

This work analysed the dynamic behaviour of a steel-concrete composite floor spanning 40m by 40m used for human rhythmic activities. The proposed numerical model adopted the usual mesh refinement techniques present in finite element method simulations, based on the ANSYS program. This model enabled a complete dynamic evaluation of the investigated floor in terms of human comfort and its associated vibration serviceability limit states.

The modal analysis performed indicated that the first twenty-four natural frequencies of the floor lie in the frequency range of aerobics. Consequently, this condition may lead the composite floor to a resonant condition. The dynamic analysis carried out showed that the accelerations values obtained when sixty-four persons practising aerobics were higher than the human comfort criteria according to five different human rhythmic force functions, namely, ISO10137, CEB, SCI, AISC and FAISCA. This situation led to human discomfort for the individuals practising aerobics on the investigated composite floor.

Considering that account that the mathematical representation of human rhythmic activities consists on a very complex problem, most of these force functions does not agree to each other. Consequently, the choice of one or other force function may lead the floor design to a more or less conservative assessment regarding the human comfort serviceability.

6 ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support for this work provided by the Brazilian Science Foundation’s CAPES, CNPq and FAPERJ.

REFERENCES


SHEAR WALL MOCK UP SUBJECTED TO MONOTONIC LOADING 
WITH CONCRETE STEEL BOND MODEL

CHETRA MANG¹, LUDOVIC JASON² AND LUC DAVENNE³

¹ CEA, DEN, DANS, DM2S, SEMT, LM2S
F-91191 Gif sur Yvette, France
and
LEME, Université Paris Ouest
F-92410 Ville d’Avray, France
chetra.mang@cea.fr

² CEA, DEN, DANS, DM2S, SEMT, LM2S
F-91191 Gif sur Yvette, France
and
IMSIA, UMR CNRS-EDF-CEA-ENSTA 9219
F-91762 Palaiseau, France
ludovic.jason@cea.fr

³ LEME, Université Paris Ouest
F-92410 Ville d’Avray, France
luc.davenne@u-paris10.fr

Key words: Reinforced concrete, Bond slip, Shear wall mock up, Crack opening

Abstract. This paper presents a numerical simulation of a shear-wall mock-up subjected to monotonic loading with a new steel-concrete bond model. First, global responses (force-displacement curves) are compared to the experimental results and show a good agreement between simulation and experiment. In order to compare the local responses such as crack orientation, crack spacing and crack opening, a new post-processing method is presented. It is based on the definition of the crack path from the displacement field or the change in the sign of the bond slip. The crack opening is then obtained using the displacement in the normal direction of the crack. The method is applied on the shear wall and the results are compared to the experimental data obtained with digital image correlation. A good agreement is once again obtained.

1 INTRODUCTION

Reinforced concrete is widely used in civil engineering structures but may be concerned by cracking especially when concrete is subjected to tensile loadings. Crack initiation and propagation are related to the interface between steel and concrete which is responsible for stress transfer, especially during cracking. Taking into account the bond slip in a numerical simulation is thus a key point when cracking wants to be correctly captured. This paper briefly introduces the bond slip model developed in [1] in the first section. Then two crack opening methods [2] are briefly described. These methods are finally applied to a shear-wall
mock-up simulated with the new bond slip model. The influences of the bond effect are especially studied.

2 DESCRIPTION OF BOND SLIP MODEL

When reinforced concrete structures are considered, one of the most classical hypotheses is to model the steel reinforcement as truss elements and to consider a perfect relation between steel and concrete. This perfect relation is applied through cinematic relation between steel and concrete, using the shape function of each element. But this hypothesis may have significant consequences when the crack properties (crack spacing and opening) are studied as the steel-concrete bond directly influences their evolution [3, 4]. To take into account the interfacial behavior, the model developed in [1], based on the previous worked from [5] is used. It is a zero thickness four node element which relates each truss element with an associated superimposed segment, perfectly bonded to the surrounding concrete (Figure 1). Each node of the interface element has three degrees of freedom (nodal displacements) (Figure 2). The relation between the generalized slip \( \{\delta(p)\} \) in the local direct frame (Figure 3) and the nodal displacements \( \{u\} \) is written in the following form:

\[
\begin{bmatrix} \delta_1(p) \\ \delta_{n_1}(p) \\ \delta_{n_2}(p) \end{bmatrix} = B(p) \{u\}
\]

with

\[
B(p) = \begin{bmatrix} \tilde{B}_1(p) & \tilde{B}_2(p) & -\tilde{B}_2(p) & -\tilde{B}_1(p) \end{bmatrix}
\]

and

\[
\tilde{B}_1(p) = 0.5(1 - p)I_3
\]
\[
\tilde{B}_2(p) = 0.5(1 + p)I_3
\]

where \( I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \) and \(-1 \leq p \leq 1\) (Figure 3)

\[\text{Figure 1: Principle of the interface element between steel and concrete [1]}\]
A constitutive law is defined between the bond stress \( \{ \sigma(p) \} = \begin{bmatrix} \sigma_t(p) \\ \sigma_{n_1}(p) \\ \sigma_{n_2}(p) \end{bmatrix} \) and the bond slip \( \{ \delta(p) \} = \begin{bmatrix} \delta_t(p) \\ \delta_{n_1}(p) \\ \delta_{n_2}(p) \end{bmatrix} \). In the tangential direction, the tangential stress \( \sigma_t \) is computed from the tangential slip \( \delta_t \) using the recommendation from [6].

\[
\sigma_t(p) = f(\delta_t(p)) \tag{5}
\]

In the normal directions, for the sake of simplicity, a linear relation is supposed between the normal stresses and the corresponding normal slips:

\[
\begin{bmatrix} \sigma_{n_1}(p) \\ \sigma_{n_2}(p) \end{bmatrix} = k_n \begin{bmatrix} \delta_{n_1}(p) \\ \delta_{n_2}(p) \end{bmatrix} \tag{6}
\]

The value of the normal stiffness is chosen high enough to be representative of a perfect bond. Additional information can be found in [1].
3 DESCRIPTION OF CRACK OPENING METHOD

A generic crack opening method is proposed to compute the crack properties (crack spacing and crack opening) of reinforced concrete structures in a complex crack pattern (crack oriented).

Before computing crack opening, crack path is determined. Two approaches are proposed for crack paths: discontinuity of displacement in the vicinity of the cracked zones or changes in sign of the bond slip. The first one is close to the experimental method developed for digital image correlation [7, 8] (see detail in [2]). The second method is more original and takes advantage of the developed bond slip model presented in the previous section. Crack position corresponds to the changes in sign of the bond slip (Figure 5). In this case, the bond slips are computed along the steel bars to locate the changes in sign. These points correspond to the crack position at the interface between steel and concrete. Compared to the first one, this method is of more direct application.

![Figure 5: Definition of the change in the tangential slip sign](image)

For both methods, from the crack positions, the crack path is determined by relating each point of a crack in its “influence” zone. This zone can be determined from the distribution of the mechanical damage for example (Figure 6).

![Figure 6: Definition of the “influence” zone from the damage distribution](image)
Crack opening is then defined as the opening in the normal direction of the crack. At each point of the crack path, the normal and the tangential crack directions \( \vec{n}_c \), \( \vec{t}_c \) and the crack angle \( \theta_c \) are computed. The normal direction is determined from the bisector of the adjacent segment of the crack path (Figure 7). The tangential direction is calculated to obtain the direct reference frame. The crack opening is then computed from the relative displacement in the normal direction at the crack position. The detail of the methodology of the crack opening can be found in [2].

![Figure 7: Definition of the crack normal direction](image)

It is to be noted that contrary to the method developed in [9, 10], the calculation of the crack opening does not imply the definition of the element length as it is directly based on the displacement and not the strain. The particularity of the method is the definition of the “post-processing line” along which the relative displacement is computed. It has to be chosen long enough to include the potential effect of regularization in the neighborhood of the crack. Hence, it becomes applicable in the case of a regularized computation ([11] for example) in which strains are not fully localized.

4 APPLICATION TO A SHEAR WALL

The new bond slip model and the new crack opening method are applied on the reinforced concrete shear-wall mock-up tested during the French national project (CEOS.FR) [12].

The dimensions of the current part of the wall are 4.2m x 1.05m x 0.15m (Figure 9). The reinforcement consists in a double steel frame (10 mm diameter) and 100 mm of space in both horizontal and vertical directions. To prevent crack opening due to bending, additional reinforcements are added at each extremity of the wall (25mm and 32mm bar). Two horizontal high strength concrete beams are connected to upper and bottom parts of the wall to ensure a correct distribution of the shear force [13].

A 3D simulation is performed using the finite element code Cast3M [15] using solid elements for concrete (30mm x 30mm x 30mm mesh) and truss elements for reinforcement (Figure 11). Steel is modeled using an elastic plastic model with linear hardening. The associated parameters are listed in table 1. Concrete, in the central part of the shear wall, is represented using a damage constitutive law developed in [16].
Figure 8: Shear wall experimental setup [14]

Figure 9: Dimensions of the shear wall

Figure 10: Reinforcement in the shear wall
Figure 11: Concrete (left) and steel (right) meshes.

Figure 12: Loading and boundary conditions

Figure 13: Bond slip – bond stress law
This law was chosen because it was successfully applied in previous work [4]. Material parameters are provided in Table 2. Top and bottom beams are considered elastic during the simulation. For the bond slip model, the bond slip law is chosen using the recommendations from [6]. The parameters are given in Table 3 and the resulting curve in Figure 13.

An increasing horizontal displacement is applied to the wall (Figure 12). A vertical force $p$ (520kN) is applied on each pre-stressed bar. A zero horizontal displacement is also applied on the right extremity of the bottom beam as in the experiment.

<table>
<thead>
<tr>
<th>Table 1: Steel parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_s$ (GPa)</td>
</tr>
<tr>
<td>190</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2: Concrete parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_c$ (GPa)</td>
</tr>
<tr>
<td>22</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3: Parameters of the bond slip – bond stress law</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bond stress (MPa)</td>
</tr>
<tr>
<td>Bond slip (mm)</td>
</tr>
<tr>
<td>Slope (Pa/m)</td>
</tr>
</tbody>
</table>

5 RESULTS AND DISCUSSION

5.1 Experimental comparison

Figure 14 illustrates the applied forces as a function of the relative displacement (difference in the horizontal displacement of the top and the bottom parts of the wall).
The global behavior is well reproduced by the simulation with especially a correct value of the maximum applied force before failure. The elastic behavior given by the sensor measurement is not totally satisfactory. It was thus checked using the relative displacement provided by the digital image correlation (Figure 14 right). In this case, the elastic evolution is correct.

The crack propagation is also correctly reproduced (Figure 15): in both simulation and experiment, an initial diagonal crack is observed at a force of 900 kN. At 2100 kN, a new crack appears on the opposite side. Finally, cracks propagate again from the loaded side.

In order to calculate the crack properties using the new crack opening method, crack paths are first determined. From the crack path method with “discontinuity” of displacement, the crack paths in the interest zone are obtained. The normal crack opening can then be computed using the methodology described in the previous section. The distribution of the normal crack opening in the zone of interest is illustrated in figure 16. From this result, the crack spacing, the mean crack opening, the maximum crack opening and the crack orientation are computed. The comparison with the experiment is shown in table 4 and 5. A good agreement is once again obtained.

**Table 4**: Crack spacing and angle

<table>
<thead>
<tr>
<th>F = 3900 kN</th>
<th>Crack angle</th>
<th>Mean crack spacing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\theta_c$ (degree)</td>
<td>(mm)</td>
</tr>
<tr>
<td>Experiment</td>
<td>28.5</td>
<td>97.4</td>
</tr>
<tr>
<td>Simulation</td>
<td>25</td>
<td>96</td>
</tr>
</tbody>
</table>

**Table 5**: Crack opening: Comparison between simulation and experiment

<table>
<thead>
<tr>
<th>F = 3900 kN</th>
<th>Mean crack opening</th>
<th>Maximum crack opening</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(µm)</td>
<td>(µm)</td>
</tr>
<tr>
<td>Experiment</td>
<td>105</td>
<td>364</td>
</tr>
<tr>
<td>Simulation</td>
<td>102</td>
<td>360</td>
</tr>
</tbody>
</table>
5.2 Influences of bond effects

In this section, the influence of the bond model is investigated by comparing to a simulation using the classical perfect relation. This so called “perfect” bond is applied using additional cinematic relations which impose the same displacement between steel and concrete (no slip).

Figure 17 provides the comparison in terms of force-displacement curves. The same evolution is globally obtained and there is no significant difference. The elastic zone is the same due to the high value of the slope in the bond stress – bond slip curve.

Figure 18 illustrates the comparison of the crack pattern (distribution of the norm of the relative displacement) at ultimate load. The crack path is not identical (differences in positions and amplitudes). A concentration at the application of the load is more significant with the perfect bond. It was expected as the perfect relation stiffens the bond between steel and concrete and may induce heavy damage.

Finally, figure 19 illustrates the distribution of the relative displacement. It shows the number of points, in percent, whose associated relative displacement is less than the indicated value. The perfect relation leads to a higher number of small crack openings (less than 200 μm) compared to the bond model. In this case, steel induces a regularization which is more significant than in the case of the bond model, especially when the damage develops. For a relative displacement equal to 900 μm, the accumulated percentage for the bond model is about 100% while it is lower for the perfect bond.
Figure 18: Comparison of the crack patterns considering the perfect relation (top) and the bond slip model (bottom) (norm of the relative displacements) at the ultimate load

It is due to the high level of degradation near the application of the load as previously reported. These differences underline the interest of taking into account the bond slip model in the simulation.

Figure 19: Evolution of the cumulative percentage of the norms of the relative displacement at the ultimate load

6 CONCLUSION

In this contribution, a shear-wall mock-up has been simulated by a new bond slip model. A comparison to the experimental results was also performed. Global responses are in good agreement. A new crack opening method was also introduced and applied to the shear-wall. A good agreement was once again achieved between experiment and simulation. Finally, the role of the steel-concrete bond has been discussed. No clear difference was observed between the simulations using the bond slip or the perfect relation. On the contrary, in terms of crack distribution, the perfect relation leads to a heavier mechanical degradation near the loading point and to a more regularized distribution in the current part, which is less in agreement with the experiment.

ACKNOWLEDGEMENTS

Partial financial support from AREVA is gratefully acknowledged.
REFERENCES


ZERO-THICKNESS INTERFACE FORMULATION FOR FRACTURE ANALYSIS OF SELF-HEALING CONCRETE

ANTONIO CAGGIANO¹, VISAR KRELANI², LIBERATO FERRARA² AND GUILLERMO ETSE¹

¹CONICET (National Scientific and Technical Research Council) and Universities of Buenos Aires and Tucuman
Buenos Aires and San Miguel de Tucuman, Argentina
acaggiano@fi.uba.ar, getse@herrera.unt.edu.ar

²Department of Civil and Environmental Engineering, Politecnico di Milano
Milano, Italia
liberato.ferrara@polimi.it, visar.krelani@polimi.it

Key words: Concrete, Self-healing, Discontinuous approach, Zero-thickness interfaces, Fracture Mechanics, Porosity

Abstract. A discontinuous-based porosity-based model for concrete subjected to time-evolution self-healing phenomena is presented in this work. The model represents an extension of a fracture energy-based elasto-plastic interface formulation which now includes porosity evolution induced by self-healing mechanisms. The formulation accounts for the characterization of concrete failure behavior in mode I and II fracture types. The post-cracking response is considered by means of specific work softening rules in terms of work spent and porosity evolution. The effects of the aforementioned phenomenon on the recovery of stiffness and load bearing capacities have been evaluated by means of three-point bending (3PB) tests performed up to controlled crack opening and up to failure, respectively, before and after conditioning. Experimental tests are employed as benchmark to validate the proposed model formulation. Particularly, after outlining the mathematical formulation of the constitutive model for interface elements, numerical analysis are compared against test data.

1 INTRODUCTION

The importance of adopting eco-friendly solutions as a fundamental requisite has become more and more important in the last years in structural engineering [1]. In this field the availability of self-healing or autogenous healing in concrete cracks, which could prevent permeation of driving factors and aggressive agents and thus controlling the structure durability, represents an important aspect for eco-friendly and green structures [2].
Consensus among the international community have been achieved about the engineering significance of concrete self-healing, which has resulted in numerous studies as well as into a clear terminology and definitions. The RILEM TC-221-SHC “Self-healing phenomena in cement based materials” distinguishes two different kinds of phenomena: i.e., (i) the “self-closing” and (ii) the “self-healing”. The first term indicates the only closure of cracks while the second one deals with the restoring of the mechanical properties. Furthermore, the RILEM TC-221-SHC proposes the difference between “autogenic” (or natural) with “autonomic” (or engineered) self-closing/healing, whether the crack closure or restoration of material properties is due to either the concrete material itself or some engineered addition [3].

In the very last decade a huge amount of researches have been dedicated to “engineered” self-healing materials, along different main directions of investigation. Particularly, the use of the so-called “crystalline additives” has been successfully explored by Sisomphon et al. [4]. The crack-healing potential of a specific two-component bio-chemical self-healing agent, embedded in porous expanded clay particles which acted as reservoir particles replacing part of regular concrete aggregates, was quantified by Wiktor and Jonkers [5]. Then, the possibility of self-healing mechanism by employing polyurethane encapsulated by glass or ceramic cylindrical capsules to seal cracks appearing at random locations in concrete beams was proposed by van Tittelboom et al. [6].

Only few models were developed in scientific literature to describe the cementitious composite self-healing. Some models were proposed to determine the amount of unhydrated cement particles in concrete specimens considering as input parameters the water-to-cement ratio and cement fineness [7]. Recently, a model aimed at simulating further hydration using water transport theory, ion diffusion theory and thermodynamics theory was proposed by Huang and Ye [8]. However, these models do not provide any information about the mechanical effects of self-healing. Then, a hydro-chemo-mechanical model was developed by Hilloulin et al. [9] to simulate the autogenous healing with the aim to explain the mechanical regains after healing processes. Di Luzio et al. incorporated in a Solidification-Microprestress-Microplane (SMM) model [10] the self-healing mechanisms through a suitably defined internal variable which accounts for the effects of cracking on diffusivity and the opposite recovering effect due to crack healing [11].

In this work, and in the framework of the discrete crack approach, an elasto-plastic porosity-based interface model for zero-thickness interfaces is proposed to predict the concrete cracking and failure behavior of concrete subject to autogenous self-healing. The interface model is formulated within the general framework of the flow theory of plasticity which is embedded in fracture mechanics concepts to account for the objectivity of the fracture energy release during post-peak regimes of concrete. Based on the original model by Carol et al. [12], the interface formulation is extended to take into account the autogenous self-healing effects in the maximum strength criterion and softening rules under mode I, II and mixed types of failure, which are differently and independently treated in the constitutive model. The relevant novel aspect in this interface model...
formulation is the inclusion of a porosity-based rule in the softening rules.

Zero-thickness interface elements, formulated in terms of contact stresses versus opening relationships, have historically been employed for modeling both the response behavior of material discontinuities such as mechanical contacts [13], bonds [14] and crack evolutions in quasi-brittle materials like concrete [15]. One of the most frequent use of interface elements in computational concrete mechanics is related to mesoscopic failure simulations. This strategy was proposed by Lopez et al. [16] for rate independent failure behavior analysis of concrete and by Lorefice et al. [17] for time dependent simulations. Recently, the aforementioned meso-scale approach was extended by the authors [18] to model the behavior of mortar-to-mortar interfaces of fiber reinforced cementitious composites.

After this introduction, Sections 2-4 highlight the formulation of the proposed porosity-based interface model for simulating the observed cracking and post-cracking behavior of concrete specimens before and after self-healing. Comparisons between experimental data and numerical predictions are presented and discussed in Section 5. Finally, Section 6 remarks the key results of the present research and figures out its possible future steps.

2 ZERO-THICKNESS INTERFACE MODEL: GENERAL ASSUMPTIONS

This section summarizes the porosity dependent interface model formulated for analyzing the fracture behavior of quasi-brittle materials like concrete after self-healing processes. The model, based on a modification of the original proposal for interface elements by Carol et al. [12], is characterized by the following constitutive equations written in rate format

\[ \dot{\mathbf{u}} = \dot{\mathbf{u}}^e + \dot{\mathbf{u}}^p \]
\[ \dot{\mathbf{u}}^e = C^{-1} \cdot \dot{\mathbf{t}} \]
\[ \dot{\mathbf{t}} = C \cdot (\dot{\mathbf{u}} - \dot{\mathbf{u}}^p) \]

(1)

where \( \dot{\mathbf{u}} = [\dot{u}, \dot{v}]^t \) is the rate of the relative joint displacement vector decomposed into the elastic and plastic components, \( \dot{\mathbf{u}}^e \) and \( \dot{\mathbf{u}}^p \), respectively. While, \( C \) defines the normal/tangential elastic stiffness matrix,

\[ C = \begin{pmatrix} k_N & 0 \\ 0 & k_T \end{pmatrix}. \]

(2)

The incremental stress vector is defined in the interface coordinates \( \dot{\mathbf{t}} = [\dot{\sigma}_N, \dot{\sigma}_T]^t \), being \( \sigma_N \) and \( \sigma_T \) the normal and shear components, respectively.

In Eq. (1) the vector of the plastic displacement rate is defined according to the following non-associated flow rule

\[ \dot{\mathbf{u}}^p = \lambda \mathbf{m} \]

(3)

where \( \lambda \) is the non-negative plastic multiplier derived from the classical Kuhn-Tucker loading/unloading and consistency conditions

\[ \lambda \geq 0, \quad f \leq 0, \quad \dot{\lambda} f = 0 \quad \text{Kuhn – Tucker} \]
\[ \dot{f} = 0 \quad \text{Consistency}. \]

(4)
In the above equations \( f = f[\sigma_N, \sigma_T] \) outlines the yield condition of the model defined by means of the following three-parameter criterion (outlining the hyperbola represented in Figure 1)

\[
f = \sigma_T^2 - (c - \sigma_N \tan \phi)^2 + (c - \chi \tan \phi)^2.
\] (5)

The tensile strength \( \chi \) (vertex of the hyperbola), the cohesion \( c \) and the frictional angle \( \phi \) are material parameters needed for identifying the inelastic state of the interface model.

Eq. (3) describes the general non-associated flow rule which mainly controls the direction \( m \) of interface fracture displacements. Particularly, the non-associated plastic direction is described by means of the transformation matrix operator \( A \) as follows

\[
m = A \cdot n
\] (6)

where

\[
n = \frac{\partial f}{\partial t} = \begin{bmatrix} \frac{\partial f}{\partial \sigma_N} & \frac{\partial f}{\partial \sigma_T} \end{bmatrix}^t
\] (7)

the associated flow vector. Then, the transformation matrix \( A \) in Eq. (6) has the following expressions

\[
A = \begin{cases} 
\begin{pmatrix} 
\tan \beta & 0 \\
\tan \phi & 1 
\end{pmatrix} & \text{if } \sigma_N \geq 0 \\
\begin{pmatrix} 
1 - \frac{\sigma_N}{\sigma_{\text{dil}}} & \tan \beta & 0 \\
\frac{\sigma_N}{\sigma_{\text{dil}}} & \tan \phi & 1 \\
0 & 0 & 1 
\end{pmatrix} & \text{if } -\sigma_{\text{dil}} \leq \sigma_N < 0 \\
\begin{pmatrix} 
0 & 0 & 1 
\end{pmatrix} & \text{if } \sigma_N < -\sigma_{\text{dil}} 
\end{cases}
\] (8)
where $\beta$ is the dilation angle of the plastic potential as highlighted in Figure 1 where $0 \leq \tan \beta \leq \tan \phi$. Thereby, the parameter $\sigma_{dil}$ represents the normal stress at which the dilatancy vanishes [19].

3 FRACTURE AND POROSITY-BASED SOFTENING RULES

The following unified function has been considered to account the softening rule of all internal parameters (controlling the evolution of the yield criterion in Eq. 5) under fracture processes which incorporates self-healing effects during the time

$$p_i = (1 - (1 - r_{pi}) S[\xi_{pi}]) SH[\psi]p_{0i}$$

where $p_i$ alternatively equals $\chi$, $c$ or $\tan \phi$, being $p_i = p_{0i}$ their initial values and $r_{pi}$ the residual parameters [20].

3.1 Fracture-based scaling function

The proposed scaling function $S[\xi_{pi}]$ is expressed as follows

$$S[\xi_{pi}] = \frac{e^{-\alpha_{pi}\xi_{pi}}}{1 + (e^{-\alpha_{pi}} - 1)\xi_{pi}}$$

based on the non-dimensional variable $\xi_{pi}$ which introduces the influence of the ratio between the current fracture work spent and the available fracture energies in mode I and II, $G_{if}$ and $G_{if}^{IIa}$ respectively,

$$\xi_{\chi} = \begin{cases} \frac{1}{2} \left[ 1 - \cos \left( \frac{\pi w_{cr}}{G_{if}} \right) \right] & \text{if } w_{cr} \leq G_{if}^I \\ \frac{1}{2} & \text{otherwise} \end{cases}$$

$$\xi_{c} = \xi_{\tan \phi} = \begin{cases} \frac{1}{2} \left[ 1 - \cos \left( \frac{\pi w_{cr}}{G_{if}^{IIa}} \right) \right] & \text{if } w_{cr} \leq G_{if}^{IIa} \\ \frac{1}{2} & \text{otherwise} \end{cases}$$

according to the $C^1$ continuity function proposed by Caballero et al. [21], while the parameter $\alpha_{pi}$ controls the decay form of such a scaling rule as shown in Figure 2.

The fracture work spent $w_{cr}$ during the opening-sliding fracture process controls the actual evolutions of the material parameters $\chi$, $c$ and $\tan \phi$ in softening regime of the interface constitutive law. The variable $w_{cr}$, defining the necessary amount of released fracture work to open a single crack in tensile and/or shear fracture mode due to normal $\sigma_N$ and/or tangential $\sigma_T$ joint stresses, has been reported in [19].

3.2 Porosity-based scaling function

The proposed porosity-based rule for the self-healing description, namely $SH[\psi]$ in Eq. (9), is expressed as

$$SH[\psi] = 1 + (1 - P_{pi}[\psi])sh$$

557
where \( sh \) (greater or equal than 0) is the self-healing scalar factor which interprets the recovery of the interface material properties due to the self-healing process. It is function of the porosity of the considered concrete through the following expression

\[
P_{p_i} [\psi] = \frac{e^{-\theta_{p_i} \psi}}{1 + (e^{-\theta_{p_i}} - 1)\psi}
\]  

(14)

where the \( \psi \) variable highlights the influence of the effective porosity as follows

\[
\psi = \frac{1 - \cos \left( \frac{\pi (\phi_{c} - \phi_0)}{\phi_f - \phi_0} \right)}{2}
\]  

(15)

being \( \phi_c \) the effective porosity \([22]\) in the interface plane, while \( \phi_0 \) and \( \phi_f \) are two input parameters to be calibrated; \( \phi_0 \) represents the maximum reachable value of the effective porosity in the concrete matrix: e.g., concrete, ideally without voids, can be represented with \( \phi_0 = 1 \). \( \phi_f \) represents that value of the effective porosity at which the interface is totally softened: concrete with 100% of voids deals with \( \phi_f = 0 \). In Eq. (14) the parameter \( \theta_{p_i} \) accounts several possible shapes of the porosity description as highlighted in Figure 3.

4 TANGENT STIFFNESS OPERATOR

The interface elasto-plastic rate equations are obtained starting from the consistency condition of Eq. (4) which expands into

\[
\frac{\partial f}{\partial t} \dot{t} + \frac{\partial f}{\partial \lambda} \dot{\lambda} + \frac{\partial f}{\partial \phi} \dot{\phi}_c = 0
\]  

(16)
where the softening parameter \( \bar{H} = -\frac{\partial f}{\partial \lambda} \) due to mechanical effects can be derived as

\[
\bar{H} = -\frac{\partial f}{\partial \lambda} = -\frac{\partial f}{\partial \rho_i} \frac{\partial \lambda}{\partial w_{cr}} \frac{\partial u^e}{\partial \rho_i} m
\]  

while the porosity evolution is taken into account through the softening parameter \( I \)

\[
I = -\frac{\partial f}{\partial \phi_c} = -\frac{\partial f}{\partial \rho_i} \frac{\partial \phi_c}{\partial \rho_i} \phi_c.
\]  

The rate of the plastic multiplier follows from the combination between Eqs. (16), (1), (17) and (18)

\[
\dot{\lambda} = \frac{\partial f}{\partial \lambda} \frac{C \cdot \dot{u}}{\bar{H} + \frac{\partial f}{\partial \lambda} C \cdot m} - I \frac{\partial f}{\partial \phi_c} \phi_c.
\]  

Consequently, the rate of normal and tangential interface stresses are derived through the rates of relative displacement vector and porosity rise as

\[
\dot{t} = \dot{C} \cdot \dot{u} + \dot{t}^\sigma \dot{\phi_c}.
\]  

Then, the tangential interface stiffness for elastic degradation expands into

\[
C^e = \left[ C - \frac{C \cdot m \otimes n \cdot C}{\bar{H} + n \cdot C \cdot m} \right]
\]  

while the rate vector of interface stresses strictly due to the rate of porosity field are

\[
\dot{t}^\sigma \dot{\phi_c} = I \frac{\partial \phi_c}{\partial \phi_c} \frac{\dot{\phi_c}}{\bar{H} + n \cdot C \cdot m} C \cdot m.
\]
5 VALIDATION OF THE PROPOSED INTERFACE MODEL

This section aims at demonstrating the validation of the proposed interface model. For the calibration purpose, experimental results on concrete specimens tested under three-point bending and presented in the work of Ferrara et al. [23] are considered.

The tests here analyzed deal with exposure under climate chamber of temperature cycles representative of typical either winter or summer daily excursions in Northern Italy. Particularly, the cycles refer to 2 weeks of accelerated climate chamber conditions. Thus, in order to evaluate the self healing capacity of concrete and its effects on the recovery of mechanical properties, at the beginning of the aforementioned curing period, the beam specimens were pre-cracked, up to different levels of residual crack opening. Specimens were pre-cracked employing a three-point bending (3pb) test set-up up to certain crack opening displacement (COD) at mid-span as a control variable.

Figure 4 shows the geometry of the considered 100 × 50 × 450 mm$^3$ concrete specimens while Figure 5 proposes the FE discretization employed in the present analysis. Plane stress hypothesis and displacement-based control are assumed. Four-node iso-parametric elements, equipped with a linear elastic model, have been adopted in the FE mesh, while all non-linearities are concentrated within zero-thickness interfaces defined throughout the adjacent edges of the finite elements in the notched zone of the beam. Non-linear porosity and fracture-based laws were introduced in those interface elements according to the formulation outlined in Sections 2-4.

The key geometric and material properties were chosen according to the experimental evidences [23]. Based on the calibration procedure, the elastic modulus and Poisson’s ratio of concrete was $E_c = 21.7 \text{GPa}$ and $\nu = 0.17$, respectively. Then, the mechanical parameters of the interface model are here listed: $k_N = 500 \text{MPa/mm}$, $k_T = 200 \text{MPa/mm}$, $\tan \phi_0 = 0.6$, $\tan \beta = 0.3$, $\tau_{\text{tan} \phi} = 0.67$, $\chi_0 = 2.5 \text{MPa}$, $c_0 = 5.0 \text{MPa}$, $G_{I}^f = 0.065 \text{N/mm}$, $G_{II}^f = 0.650 \text{N/mm}$, $\sigma_{\text{dil}} = 10 \text{MPa}$, $\alpha_\chi = 0.5$, $\phi_c = 0.46$, $\phi_0 = 0.80$, $\phi_f = 0.20$, $sh = 0.3$ and all remaining parameters were considered null.

As example, Figure 6 shows the force-deflection curves against the corresponding experimental results. It can be observed that the post-cracking response is well captured through the considered discontinuous approach based on non-linear interfaces. Furthermore, the comparison between the experimental and numerical results in terms of the
load vs. crack opening displacement (COD) highlights as the model is able to predict the self-healing recovery of the load bearing capacity with respect to the unloading value at which the value decayed in the pre-cracking stage.

A recovery of the flexural load equals to about 250N is well reproduced by the numerical model, which corresponds to a recovery of about 17.05% of the maximum load capacity. However, and due to re-hydration self-healing phenomenon, a significant increase of the fracture energy takes place in the second-stage (post-treatment) of results of the analyzed three-point bending tests. This is introduced in the model throughout tabulated data which allows to give an opportune variation of the fracture energy against the porosity evolution. Particularly, based on the calibration data obtained from the numerical results, an increment of 12.07% of fracture energy in mode I has to be considered for the proposed analyses presented in this section.

6 CONCLUDING REMARKS

This paper addressed the formulation of a fracture and porosity-based interface model for analyzing the self-healing capacity of cementitious composites. Particularly, the proposed formulation explicitly modeled the porosity evolutions due to self-healing pore closures through functions which affect the strength parameters and softening rules of the proposed discontinuous model. Numerical analyses demonstrated the predictive capabilities of the constitutive model in terms of the most relevant aspects of the mechanical behavior of concrete after crack closure due to self-healing processes. The proposed interface model can be employed in mesoscopic analyses using mortar-mortar and mortar-aggregate interfaces aimed at simulating failure processes of concrete specimens under self-healing phenomenon. This aspect is beyond the scopes of the present paper and will be addressed in future developments of the current research.
Figure 6: Load-deflection behavior on three-point beam: experimental results [23] vs. numerical predictions.

ACKNOWLEDGMENT

The authors acknowledge the financial support for this work obtained by CONICET (Argentinian National Council for Science and Technology) through the Grant PIP 112-200801-00707, by CIUNT (Research council of the National University of Tucuman) through the Grant 26/E479 and by the EnCoRe Project (FP7-PEOPLE-2011-IRSES n. 295283; www.encore-fp7.unisa.it) funded by the European Union within the Seventh Framework Programme.

REFERENCES


1 INTRODUCTION

Frictional contact analysis is traditionally considered among the most challenging problems of computational solid mechanics. Firstly, the analysis is geometrically non-linear with potential convergence issues due to deformation-driven boundary conditions. Secondly, frictional definitions provide upper boundaries of sliding resistance rather than fixed relationships between contact local stresses. It may be potentially reflected in multiple solutions satisfying, nevertheless, all required equilibrium and deformation conditions. This potential non-uniqueness may be associated, therefore, with poor convergence, erroneous results, etc. These challenges can be especially important in case of high friction, i.e., for problems where friction-based responses (e.g., traction, wear, abrasion) are of prime interest of the analysis and cannot be ignored or simplified.

Another important issue is the coupling nature of the input/output in the analysis. Indeed, shear frictional stresses are expected to be the output of the analysis. However, this output may control the entire stress distributions including normal contact stresses, i.e., needs to be known in one form or another. Although this coupling interaction can be potentially handled by different means, a simple iterative approach can be especially helpful in computational implementations. According to the approach, initially introduced unknown (“coarse”) stresses $\tau$ are used as an integrated part of applied load to predict normal contact stresses $\sigma$. Then updated (“refined”) stresses $\tau'$ are calculated as functions of known normal stresses $\sigma$, and such iterative process continues until desired accuracy is achieved:
Here, $\mu$ is the coefficient of friction (COF) which can be either constant or a function of the local stress-strain state; and $\Phi$ is the correlation function between the “coarse” and “refined” solutions. The iterative process schematically outlined by (1) can potentially include a) 1-, 2- and 3-D statements; b) static, dynamic and transient definitions; c) quite complex frictional laws accounting for effects of local deformation, velocity. In addition, it can provide a very convenient qualitative understanding of the deformation process and its correlation with intuitive expectations. Major advantage, however, is in relative simplicity of computational implementations in connections with either in-house or commercial general purpose FEA codes. This iterative concept was successfully demonstrated, for example, in analysis of fiber/matrix interfacial crack growth [1] and its computational efficiency was shown, for example, in Ref. [2] for integration with ABAQUS.

In spite of advantages of iterative approaches (1), their applicability is not unlimited, especially, in cases of severe friction [3]. It was shown and explained [3] that there is a set of conditions where direct implementation of concept (1) cannot provide convergent solutions. Based on understanding of such limitations, more robust iterative schemes were proposed [3], for example, in the following form

$$ \tau \rightarrow \sigma(\tau) \rightarrow \tau' = \tau - z[\tau - \Phi(\tau)] = \Psi(\tau) $$

(2)

where scalar parameter $z$ can be selected to mitigate limitations of convergence and guarantee, at the same time, the “correct” solution. Successful implementation of concept (2) for one-dimensional and one-element frictional problems was shown in Ref. [3]. The present study is focused on expansion of this enhancement to more complex multi-element (and/or multi-dimensional) problems including definitions of convergence criteria, techniques to accelerate the convergence rate, associated numerical challenges and proposed benchmarking problems for assessment of computational robustness.

2 DEMONSTRATION CASE

To illustrate potential challenges of frictional analysis consider a simple demonstration case with well-known closed-form solution used as a reference of computational accuracy. A suggested example is shown at Figure 1a, and it represents a tension element (e.g., rope) with multiple wraps around a cylindrical body with radius $r$ ($\alpha$ is the total wrap angle; and $b$ is the rope width). There is friction between the rope and the body defined by constant COF=$\mu$.

Thus, if tension $T$ is applied at the beginning of the rope, corresponding equilibrium tension at the end $R$, will be smaller due to frictional resistance. Assuming that the COF is independent of contact pressure, slip velocity, etc., the classical Euler’s solution is directly applicable to this problem. According to the Euler’s solution (Figure 1b), tension reduction $dT$ is compensated by frictional stresses $\tau$ as

$$ dT(\alpha) = \tau(br)d\alpha = (\sigma \mu)(br)d\alpha = \left( \frac{T}{br} \mu \right)(br)d\alpha = T(\alpha)\mu d\alpha $$

(3)

with a simple solution for obvious boundary conditions $T(0)=T$ and $T(\alpha)=R$ as.
\[ R/T = \exp(-\mu \alpha) \tag{4} \]

It is easy to see that this solution does not possess any restrictions with respect to the input data (either the COF or the wrap angle) and should provide a unique quantitative assessment for this frictional problem. In other words, any robust numerical analysis is expected to generate similar results within specified margins of computational accuracy.

For numerical implementation of this problem, consider its discrete definition schematically shown at Figure 1c. In this case, the rope is considered as a combination of \( N \) elements, and each \( k \)-th element is under local tensions \( T_{k-1} \) and \( T_k \), respectively (Figure 1d). Tension reduction \( \Delta T_k \) can be calculated similarly to (3) as

\[ \Delta T_k = \sigma_k \mu \beta r \Delta \alpha = \left( \frac{T_k + T_{k-1}}{2} \right) \mu \Delta \alpha \tag{5} \]

where \( \Delta \alpha = \alpha / N \). Since distribution of tensions \( T_k \) is not known \( a \ priori \), the following equilibrium condition should be applied (Figure 1d):

\[ T_k = T_{k-1} - \Delta T_k \tag{6} \]

If the “coarse” solution \( T_k \) is known, the “refined” solution \( T'_k \) can be calculated therefore in the next iteration as:

\[ \begin{bmatrix} T' \end{bmatrix} = [A] \begin{bmatrix} T \end{bmatrix} + [B] \tag{7} \]

where

\[ [A] = \begin{bmatrix} -w & 0 & 0 & \ldots & 0 & 0 \\ v & -w & 0 & \ldots & 0 & 0 \\ 0 & v & -w & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & -w & 0 \\ 0 & 0 & 0 & \ldots & v & -w \end{bmatrix}; \quad [B] = \begin{bmatrix} T_v \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \tag{8} \]

\[ w = \mu \Delta \alpha / 2; \quad v = 1 - w \tag{9} \]

The iterative process (7) continues until desired convergence with respect to the reference solution (4) is achieved.

Assuming the initial solution \( T_k = T \) (iteration “0”), Figures 2 and 3 illustrate convergence process (7) as functions of COF and \( \alpha \), respectively, at \( N = 10 \) and \( T = 1 \). In both parametric studies, there are scenarios with very quick and robust convergence (Figures 2a, 3a) and relatively slow or even non-converged solutions (Figures 2b, 3b). Corresponding comparisons with the closed-form solution (4) are also shown in Table 1 for representative convergent cases. These results (Table 1) indicate that, in case of convergence, the
computational predictions are practically identical to the “exact” expectations, i.e., iterative scheme (7) works correctly in principle.

Table 1 Comparison between numerical and closed-form solutions

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$\mu$</th>
<th>Iteration 10</th>
<th>Iteration 20</th>
<th>Iteration 30</th>
<th>Solution (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$2\pi$</td>
<td>0.1</td>
<td>0.52261</td>
<td>0.53338</td>
<td>0.53338</td>
<td>0.53349</td>
</tr>
<tr>
<td>$2\pi$</td>
<td>0.2</td>
<td>0.26109</td>
<td>0.28414</td>
<td>0.28414</td>
<td>0.28461</td>
</tr>
<tr>
<td>$2\pi$</td>
<td>0.4</td>
<td>0.05532</td>
<td>0.07993</td>
<td>0.07993</td>
<td>0.81003</td>
</tr>
<tr>
<td>$2\pi$</td>
<td>0.6</td>
<td>0.00881</td>
<td>0.02196</td>
<td>0.02203</td>
<td>0.02305</td>
</tr>
<tr>
<td>$0.5\pi$</td>
<td>1.0</td>
<td>0.18108</td>
<td>0.20721</td>
<td>0.20721</td>
<td>0.20788</td>
</tr>
<tr>
<td>$\pi$</td>
<td>1.0</td>
<td>0.02303</td>
<td>0.04208</td>
<td>0.04210</td>
<td>0.04321</td>
</tr>
<tr>
<td>$1.5\pi$</td>
<td>1.0</td>
<td>0.00171</td>
<td>0.00771</td>
<td>0.00821</td>
<td>0.00898</td>
</tr>
</tbody>
</table>

3 CONVERGENCE CRITERIA

Although iterative solutions provide, in general, quick results and relatively simple computational implementations, their application is not un-conditional. To formalize limitations of the iterative processes, governing equations (7) are defined in a more general potentially non-linear form between the coarse ([X]) and the refined ([X']) solutions as

$$[X'] = [\Phi(X)]$$  \hspace{1cm} (11)

In this case, according to the Fixed-Point method [4], requirements of convergence for all $k = 1, \ldots, N$ are

$$|\lambda_k| < 1$$  \hspace{1cm} (12)

where $\lambda_k$ are eigenvalues of Jacobian matrix $[J_\Phi]$ defined for matrix $[\Phi]$. At $N=1$, condition (12) is simply reduced to [4]

$$|\partial\Phi_i / \partial X_j| < 1$$  \hspace{1cm} (13)

as was demonstrated [3] for one-dimensional frictional problems.

For the considered demonstration case, general criteria (12) provide the following simple assessments:

$$[J_\Phi] = [A]; \quad \lambda_k = w; \quad w = (\alpha \mu)/(2N) < 1$$  \hspace{1cm} (14)

Indeed, non-convergent solution was shown at COF = 3.183 ($N=10; \alpha = 2\pi$), i.e., at $w=1$ (Figure 2b). Similarly, another limit of convergence providing $w=1$ was shown at $\alpha = 1,146$ deg at $N=10$; COF=1 (Figure 3b). These examples demonstrate applicability of the convergence criteria (12) to iterative solutions of frictional problems.

4 ENHANCEMENT OF CONVERGENCE

Criteria of convergence (12) provide a powerful mechanism to improve robustness of computational implementation. Consider first a case where conditions (12) are satisfied. The
The easiest way to accelerate the convergence would be to modify original iterative statement (11) in form of

$$[X'] = \left[\Psi([X])\right]$$

(15)

where

$$\left[\Psi([X])\right] = [X] - z\left[[\Phi(X)] - [X]\right]$$

(16)

i.e., the iterative process is substituted by the following updated statement:

$$[X'] = [X] - z\left[[\Phi(X)] - [X]\right]$$

(17)

Note that at $z = -1$, process (17) is reduced to the original form (11).

Parameter $z$ should be selected to satisfy condition (12) for the updated Jacobian as

$$\begin{bmatrix} [J_{\Psi}] = [I] - z\left[[J_{\Phi}] - [I]\right] = (1+z)[I] - z[J_{\Phi}] \end{bmatrix}$$

(18)

where $[I]$ is the identity matrix. This condition for the demonstration case provides the following simple assessments:

$$-2/(1+w) < z < 0$$

(19)

Figure 4 shows the effect of parameter $z$ on convergence of the demonstration case at $\mu = 1$; $\alpha = \pi$; and $N = 10$, where parameter $z$ is defined within boundaries (19), i.e., between -1.728 and 0. It is easy to see that the fastest convergence in 10 iterations is achieved at $z^* = -0.864$ (Figure 4a). This value corresponds to a case where eigenvalues of introduced matrix (18) are zero:

$$z^* = 1/(1+w)$$

(20)

Note that, at $z \neq z^*$ but within condition (19), the convergence still exists but only requires more iterations to achieve desired accuracy (Figure 4b). In all cases shown at Figure 4, the results converge to the closed-form expectations (4).

If a unique value of parameter $z$ is not applicable (e.g., diagonal elements of matrix $[J_{\Phi}]$ are of different signs), a more complex definition can be suggested in form of

$$[X'] = [X] - z\left[[Z][\Phi(X)] - [X]\right]$$

(21)

where $[Z]$ is either a diagonal matrix with different diagonal elements or, in a general case, a full non-diagonal matrix with all (or majority) of non-zero elements. In these cases, the fastest convergence is expected when eigenvalues of Jacobian for (21) are zero, i.e., at

$$[Z^*] = \left[[J_{\Phi}] - [I]\right]^{-1}$$

(22)

If conditions (12) are not satisfied, introduction of enhancement modifications (17) or (21) still can help to achieve convergence. However, the results of convergence need to be checked if they are the correct ones, i.e., if they correspond to the physical expectations. The easiest way to do it is to change parameters of the problem, for example, the number of elements. If the results are different, more refined meshes can be suggested.
5 APPLICATIONS

In addition to contact problems with severe constant COF, the considered enhancement and boundaries of successful convergence may be especially helpful in other applications as well. One of them is a case where COF itself is a function of local deformation, e.g., contact pressure or sliding velocity (see, for example, Ref. [5]). Due to potential sensitivity of local analysis to computational implementation (e.g., mesh density), applications of enhancement techniques (17) or (21) can be helpful to achieve desired results. Another area of potential application is probabilistic modeling of frictional contact, where statistical simulation (Monte-Carlo) technique or more advanced methods (FORM/SORM, sampling, etc.) are used. In these cases, extreme realization beyond conditions (12) can be randomly generated with corresponding risks for the entire analysis. The risk can be especially high if rare-probability effects are of prime interest.

Although this work addresses contact analysis only, similar convergence issues can be potentially expected in other statements of computational mechanics where iterative processes are applied. Among relevant problems are materials with complex constitutive laws of deformations sensitive, for example, to deformation rate, progressive damage, plasticity, etc. In spite of difference in the physical natures of considered problems, their computational limitations could be quite similar to (12)-(13).

In addition, the demonstration problem considered in this work (Figure 1) can be suggested as a frictional benchmarking problem for robustness evaluation of FEA solutions. Indication of the robustness would be generation of convergent solution at $w=1$ without modification of the mesh density. The solution should be then compared with the “exact” result (4) to make final conclusions regarding the computational efficiency.

6 CONCLUSIONS

It is shown that there is a class of computational frictional problems where convergent solutions cannot be achieved without modifications of modeling parameters (e.g., mesh density) or adjustments of iterative processes. General criteria for such problems are defined. Enhancement techniques are suggested for acceleration of modeling convergence. Suggested criteria and enhancements are demonstrated on a representative example with known closed-form solution. The demonstration example can be used as a benchmarking problem for assessment of robustness of existing and future frictional modeling approaches. It is also noted that the similar nature of convergence issues can be expected in other problems of computational mechanics applying iterative concepts of analysis.

7 ACKNOWLEDGEMENTS

The author thanks the United Technologies Research Center (UTRC) for permission to publish this work and Dr. Sergey Shishkin (UTRC) for his comments and suggestions.

REFERENCES


---

Figure 1. Scheme of the demonstration example.
Figure 2. Effect of coefficient of friction on convergence.
Figure 3. Effect of contact angle on convergence.
Figure 4. Sensitivity of convergence to the enhancement parameter.
A NEW FORM OF FINITE ELEMENT METHOD FOR PLANE STRESS/PLANE STRAIN PROBLEMS IN REINFORCED CONCRETE CONSIDERING MATERIAL NONLINEARITY AND CRACKING

KONSTANTIN G. ASHKINADZE

¹ Jacobs Engineering Ltd.
#1800, 10065 Jasper Avenue
Edmonton, Alberta, Canada
konstantin.ashkinadze@jacobs.com
www.k-the-eng.com

Key words: Reinforced Concrete, Finite Element Analysis, Nonlinearity, Cracking, Plane Stress/Plane Strain, Ultimate Capacity Analysis.

Abstract. This paper proposes a new numerical analysis procedure for reinforced concrete structures with material nonlinearity and cracking. Traditional finite element analysis (FEA), with its reliance on matrix algebra, is an essentially linear procedure. To use it for nonlinear problems, the loading has to be performed in small steps and iterative procedures used to distribute residual nodal forces on each step. The convergence of these procedures degrades after the events of fracturing or cracking in the concrete. If the cracking is extensive, the algorithm may become critically unstable, which may not be equivalent to physical failure.

Instead, it is proposed to model the response of reinforced concrete structures in simple tests “constant preload + active load to failure” as a non-step procedure going straight to the target load. To accomplish this, deformation properties of concrete and steel are described by non-associated work-hardening theory of plasticity, formulated such that the plastic strains have a potential. The method is formulated in terms of stresses because the strength criteria are more readily formulated in these terms and because equations of static equilibrium (Navier) are easy to solve. Material nonlinearity of concrete and steel and imperfect bond between them is accounted for by direct method of energy, which is a finite function of stress even if the deformation curve is nonlinear.

A variational principle of minimum complementary energy is formulated and a numerical method developed for minimizing it on the domain of statically admissible stress fields (SASF). For construction of SASF for arbitrarily shaped structures in the case of plane stress/plane strain problems, a variant of finite element procedure is proposed. This allows exclusion of “dependent” variables and formulation of the energy functional minimization problem in terms of “fundamental” variables only, which in 2D problems are only one per node. Cracking in the concrete is considered by imposing additional conditions on SASF.

For functional minimization, a variant of first-order variable metric method is developed to accommodate “leaps” in the path of descent due to cracking. A methodology for ultimate load capacity analysis is proposed, based on solution of a double optimization (or maximin) problem in the SASF domain. Thus, the failure load is determined as part of the normal computational process and not as a disruption of this process, which improves its credibility.
1 INTRODUCTION: LIMIT STATE ANALYSIS OF REINFORCED CONCRETE STRUCTURES

The modern state of construction industry is characterized by broader use of 2D and 3D reinforced concrete structures (plates, shells, bearing walls) besides the traditional frame structures. For determination of internal forces and stresses in these structures, in the first approximation, linear-elastic finite element (FE) analysis is used. However, linear FE methods have limited application to reinforced concrete due to its material nonlinearity (deviation of the laws of deformation of concrete and steel from linear) and structural nonlinearity (cracking, compression crushing, reinforcement bond failure).

A more advanced approach is direct account for cracking and plasticity in the FE analysis, which then becomes nonlinear. However, the global system of equations for the structure:

\[ K\Delta = P \]  

is the same as in the linear elastic case \((K = \text{global stiffness matrix}, \Delta = \text{nodal displacement vector}, P = \text{nodal force vector})\). This method is linear in nature and is poorly suited for nonlinear problems when \(K\) matrix becomes dependent on \(\Delta\) and iterative algorithms such as Newton-Raphson procedure have to be used. These procedures are very laborious and have poor convergence near the points of state change of the structure. Any minor cracking in the tension zone of the structure, which has no real effect on its ultimate capacity, critically impedes the performance of the method [1]. The load has to be broken down into small steps and convergence achieved on each step, with the locally failed zones weighing heavily on it.

The greatest problems with these methods occur on the final stages of the process of loading when nonlinearity is pronounced. The ultimate failure of the structure is usually established by critical divergence of the iterative process; it is essentially the failure of the algorithm, not of the structure. These last steps produce the least credible results; yet it is these stages that are of the primary interest in the analysis, whereas the extensive analytical data obtained on the early stages of loading is largely useless.

2 THE IDEA OF THE METHOD

The idea of the method is to avoid the need for step-by-step analysis and seek for the solution in the ultimate condition at once. This seemingly contradicts the known facts of inelasticity and load history dependence of reinforced concrete. However, the purpose of the ultimate limit state analysis is not to follow the history of response. In these types of problems, the load history usually consists of initial constant load (preload) followed by monotonically increasing active load to failure. In this simplified approach, the material is considered nonlinear-elastic; the effect of unloading and hysteresis is ignored.

The suitable technique is sought in the use of variational (energy) methods. Energy considerations are routinely used in the finite element formulations [2]; however, in the end the energy functionals are reduced to the same form as the direct stiffness formulation (Equation (1)). At the same time, directly considering the expression

\[ U = \int_V udV \rightarrow \min \]  

where
is more promising because the noted integrals can be readily calculated for any shape of the material deformation diagram, not necessarily linear. The energy of deformation is additive for concrete, reinforcement, and the contact between them, thereby providing a uniform basis for account of all components of the system. (Here and below, tensor notation is used, with repeating indices meaning summation from 1 to 3; \( \sigma = \sigma_{ii}/3 \) and \( \theta = \varepsilon_{ii} \) – hydrostatic stress and strain; \( T = (s_{ij}s_{ij})^{1/2} \) and \( \Gamma = (e_{ij}e_{ij})^{1/2} \) – deviatoric stress and strain intensity, respectively; \( s_{ij} \) and \( e_{ij} \) = components of the stress and strain deviator.)

The functional in Equation (2) is expressed via components of strain (or displacement). As seen from Equation (1), FE formulations in terms of displacements are the most common. However, we chose to formulate the model in terms of stresses, for the following reasons:

- Equations of static equilibrium can be expressly satisfied outside of the minimization, which greatly reduces the number of unknowns. For plane stress, this method has 2 times fewer fundamental variables than displacement method on the same mesh.
- Strength and cracking equations are more readily stated in terms of stresses.

3 CONSTITUTIVE MODEL FOR CONCRETE

Vast literature is dedicated to the problem of constitutive modelling of concrete in multidimensional numerical analysis of structures, however no universally applicable model exists. The constitutive model used herein was formulated to satisfy the following conditions:

- The material is considered isotropic, hence all models are formulated in terms of invariants of the corresponding tensors.
- For inelastic deformation of the concrete, the theory of plastic flow (incremental) with isotropic strain-hardening is used.
- The concrete has the property of internal friction and dilatancy, which may deviate from the normality rule [5]. This is modeled by non-associated law of plasticity, in which the plastic incremental strains have a potential but this potential is not related to the strength condition for the concrete.
- The strength condition for the concrete is taken in the Schleicher-Nadai form [3]:

\[
T^2 - \varphi(\sigma) = 0
\]  

(4)

- Two state-changing conditions are considered: (a) fracturing in case of compression and (b) cracking in case of tension. In the first case, all stress components are set to zero. In the second case, normal and tangential stresses to the crack planes are set to zero. Both changes are final (i.e. cracks and microcracks do not close).

3.1 Plastic potential

The key assumption of the model states that the plastic strains of concrete have a potential:

\[
\varepsilon_{ij}^{pl} = \frac{\partial F}{\partial \sigma_{ij}}
\]  

(5)
This condition is necessary for the plastic strain energy to be a function of the current state of the material, independent of the pre-history of loading. The existence of this potential is compatible with thermodynamics of irreversible processes. It can be shown from the model that the plastic potential \( F \) can be any function of “pseudo-time” \( \zeta \), which is not the physical time but can be any monotonically increasing parameter characterizing the length of the passed trajectory of loading. In this model, the parameter \( \zeta = \sqrt{T^2 + \sigma^2} \) = the length of the radius-vector of the current point in the space of stresses, is chosen as the “pseudo-time”.

The actual form of the function \( F(\zeta) \) can be established from a uniaxial compression test. In this stressed state, \( T = \sigma_1 \sqrt{2/3} \) and \( \sigma = \sigma_1 \cdot 1/3 \), which gives \( \zeta = \sigma_1 \cdot \sqrt{2/3} \). Composing an expression for complementary energy of plastic deformation using Equation (5):

\[
\Phi_{pl}(\sigma_{11}) = \int_0^{\sigma_{11}} d\sigma_{11} = F(\zeta) \left|_{\zeta=\sigma_1 \cdot \sqrt{2/3}} \right.
\]

(6)

On the other hand, \( \Phi_{pl}(\sigma_{11}) = \int_0^{\sigma_{11}} d\sigma_{11} = \frac{\sigma_{11}^2}{2E_0} \) is known from the \( \sigma - \epsilon \) diagram obtained in the compression test. This identifies the function \( F(\zeta) \) uniquely.

The non-associated nature of the adopted law of plastic deformation is manifested in that \( F \) is a separate function, not related to the strength condition (Equation (4)). For concrete, the use of non-associated laws of deformation is natural because these materials exhibit not only plastic but also brittle failures, hence their behaviour is described by multiple limit functions.

3.2 Dilatancy

Dilatancy is the phenomenon of volumetric expansion under the action of shear stresses only, without hydrostatic stress. However, the actual dilatancy measured in the tests is usually substantially less than that calculated by the normality rule of plastic flow [5].

The hydrostatic strain can be computed from Equation (5) as follows:

\[
\theta^{el} = \frac{F'(\zeta)}{\zeta} \sigma
\]

(7)

In reality, however, dilatancy may be a totally different function depending on both \( T \) and \( \sigma \). The resulting additional complementary energy per unit volume will be:

\[
\Phi_{dil} = \int \epsilon d\sigma + \int e_{yy} ds_{yy} = \int \left( \theta_{dil}(T, \sigma) - \frac{F'(\zeta)}{\zeta} \sigma \right) d\sigma
\]

(8)

This expression can be calculated in finite form only if

\[
\theta_{dil}(T, \sigma) - \frac{F'(\zeta)}{\zeta} \sigma = \psi(\sigma) \text{ only}.
\]

(9)

This condition leaves limited room for approximation of experimental data on dilatancy. However, it allows correction for the available data in integral form or by collocation method.
3.3 Full strain potential

The resulting full strain potential has the form:

$$\Phi = \frac{\sigma^2}{2K_0} + \frac{T^2}{2G_0} + F(\zeta) + \int \psi(\sigma) d\sigma$$

where $K_0$ = initial (linear elastic) bulk modulus, $G_0$ = linear elastic shear modulus of the concrete. The first two terms represent elastic component of strain. The total strains are

$$\varepsilon_y = \varepsilon_y^e + \varepsilon_y^d = \frac{\partial \Phi}{\partial \sigma_y}$$

4 CONSTITUTIVE MODEL FOR REINFORCEMENT

Most concrete structures are reinforced. Deformation compatibility between the steel and the concrete is assured by bond stresses along the surface of the bar or by anchoring devices at the ends of the bar. Near cracks, the strains in the rebar concentrate and cause bond slippage, essential for assessment of crack opening, rebar stresses, stiffness and strength of the structure.

The contact phenomena between the rebar and the concrete are studied by technical theory of reinforcement bond. Following [6], the resolvent equations of this theory are:

- Deformation compatibility condition:

$$\varepsilon_s - \varepsilon_c = \frac{dg}{ds}$$

where $\varepsilon_s$ = strain in the reinforcement; $\varepsilon_c$ = strain in the concrete in the direction of the rebar; $g$ = relative slippage of rebar and concrete; $s$ = coordinate measured along the bar length.

- Condition of static equilibrium of a segment of the rebar in the concrete:

$$A_s \frac{d\sigma_s}{ds} = P_s \tau_c$$

where $A_s$ = section area of the rebar; $P_s$ = its perimeter; $\sigma_s$ = axial stress in the reinforcement and $\tau_c$ = bond stress.

- The “law of bond” - a constitutive relation:

$$\tau_c = f(g)$$

Typically, $f(g)$ is a significantly nonlinear function. Different forms of “laws of bond” are described in [6]. This expression can be inverted to the form $g = f^{-1}(\tau_c)$.

The complementary energy of deformation per unit volume of the rebar is:

$$\Phi_{sk} = \int \varepsilon_s d\sigma_s ; \quad \varepsilon_s = \frac{d\Phi_{sk}}{d\sigma_s}$$
4 VARIATIONAL PRINCIPLE OF REINFORCED CONCRETE DEFORMATION

Equations (10), (15) and (16) express the complementary energy of the system in terms of stresses. Now we shall formulate a variational principle stating that the functional

$$ J = \int \int \Phi dV + \sum_{k=1}^{r} \left( A_s^k \int \Phi_{sk} ds + P_s^k \int \Phi_{sk} ds \right) - \int \int u_i^* \sigma_{ij}^c n_j dS $$

(17)

being considered on the class of all statically admissible stress fields, has a minimum value on the actual stress field. In Equation (17), \( r \) = the total number of reinforcing bars in the structure; \( S_u \) = part of the boundary on which the displacements, \( u_i^* \), are given; \( n_j \) = the exterior normal vector to the boundary; \( \sigma_{ij}^c \) is the tensor of stresses in the concrete only. Equation (17) is a generalization of Castigliano’s variational principle [4].

The “statically admissible” stress field, by definition, is one that satisfies the equations of equilibrium (Navier) in the volume; static boundary conditions on the part of the boundary \( S_T \) where external loads are applied; and static equilibrium equations for each rebar in the form of Equation (13). The Navier’s equations are to be written in terms of “total” stresses:

$$ \sigma_{ij}^{tot} = 0 ; \sigma_{ij}^{tot} + \sum_{k=1}^{r} A_i^k n_i^k n_j^k \sigma_{ij}^c \int \Delta(\zeta - s) ds $$

(18)

where the index separated by a comma means differentiation by this index; \( \Delta(\zeta - s) \) is three-dimensional Dirac’s delta function and \( n_i^k \) are the directive cosines characterizing the direction of the \( k \)-th rebar. Tensor \( \sigma_{ij}^{tot} \) is a symmetrical tensor of the same dimension as \( \sigma_{ij}^c \).

The static boundary conditions on the part of the boundary \( S_T \) can also be formulated in terms of total stresses, as follows:

$$ \sigma_{ij}^{tot} n_j = T_i^* $$

(19)

\( T_i^* \) = the external forces (loads) acting on the structure).

The proof of the variational principle is done in ways standard for functionals of Castigliano type [4]. It establishes that \( \delta J = 0 \) and \( \delta^2 J > 0 \) for the stress field that is not only statically admissible but also satisfies Cauchy’s strain compatibility equations and the material constitutive equations (10-11), (15) and (16), i.e. for the actual stress field.

A theorem of uniqueness of the solution can be proven for this functional, stating that the stress and displacement field in the reinforced concrete structure satisfying all static and kinematic boundary conditions, Navier’s equilibrium equations, Cauchy’s strain compatibility equations and the material constitutive equations, is defined uniquely in the elastic-plastic parts of the system. In the fractured parts of the system, the theorem of solution uniqueness is invalid; strains and displacements in these areas are indefinite.
5 THE FINITE ELEMENT DISCRETIZATION

Up to this point, the analysis was fully generic and 3D. From now on, we will focus on plane-stressed structures. As the first step, the complementary energy functional needs to be discretized in the space of stresses. Introducing a mesh of finite elements, can write:

$$\sigma_{ij}(\xi) = \sum_{l=1}^{n} \sigma_{ij}^l \varphi_l(\xi)$$

(20)

where $\varphi_l(\xi)$ is the finite element shape function, equalling 1 at the $l$-th node and 0 in all other nodes; $n$ is the total number of nodes and $\xi$ is the generalized coordinate.

Although the problem as a whole is nonlinear, the static equilibrium (Navier’s) equations are linear. Therefore, in construction of statically admissible stress fields, the apparatus of matrix algebra is appropriate. Discretizing Navier’s equations to an arbitrarily shaped domain by means of Equation (20), obtain a linear system of algebraic equations linking together the “fundamental” and “dependent” stress variables. In matrix notation,

$$A_f \sigma_f + A_d \sigma_d = \bar{P}$$

(21)

where indices “f” and “d” denote “fundamental” and “dependent” variables. The matrix $A_d$ must be quadratic and non-degenerate. Some of the boundary conditions (= restraints) are set in terms of displacements, not stresses, so the unknown reactions at these supports need to be also included in the $\sigma_f$ vector. When “fundamental” variables are obtained from functional minimization process, “dependent” variables can be found by reverse pass in Equation (21).

6 MINIMIZATION OF THE ENERGY FUNCTIONAL

All practical methods of minimization are classified as non-gradient (0th order), gradient (1st order) and Newtonian (2nd order), which use the matrix of second derivatives of the target function. The gradient of the subject functional $J$ can be computed in a finite form as

$$\frac{\partial J}{\partial (\sigma_{ij})} = \int_{A} \left( \frac{\partial \Phi}{\partial T} s_{ij} + \frac{\partial \Phi}{\partial \sigma} \cdot \frac{1}{3} \sigma_{ij} \right) \varphi_d dA - \int_{S} \sigma_{ij} \varphi_d dS$$

(22)

($t$ = the thickness of the structure at the current point), because the function of global complementary energy $\Phi(T, \sigma)$ is defined analytically.

Note however that the gradient must be expressed in terms of fundamental variables only. Considering Equation (21), obtain:

$$\left( J'_r \right) = \begin{bmatrix} E \\ -A_d^{-1} A_f' \end{bmatrix} \left( J'_0 \right) \equiv G \left( J'_0 \right)$$

(23)

where $\left( J'_r \right)$ = the gradient vector in terms of fundamental variables only, $\left( J'_0 \right)$ = the gradient vector in terms of all components of stresses, and $E$ is the unity matrix.

The transformation matrix $G$ is very large. Its dimensions (exclusive of the unity matrix) are $(n + n_l) \times 2n$ where $n$ = total number of nodes, $n_l$ = the number of boundary restraints. It needs to be stored because it is required on every iteration; and it can be stored efficiently
using its block structure. The corresponding algorithm has been developed, in particular, for planar structures discretizable with a regular rectangular mesh. Matrix $G$ is constant in the absence of state changes in the system (i.e. cracking), after which it needs to be modified.

In choosing the minimization method, we limited our analysis to methods not more than 1st order that have the property of finite-step convergence in the linear case, namely [7, 8, 9]:
- Conjugate gradients method (1st order);
- Variable metric method (quasi-Newton, asymptotically tending to 2nd order);
- Powell method (non-gradient);
- Dual directions method (similar to variable metric method).

All finite-step methods of quadratic programming are based on the following idea: starting from the point $\bar{u}_i$, choose a direction of descent $\bar{p}_i$ and determine the next point by the rule

$$\tilde{u}_{i+1} = \bar{u}_i - \alpha_i \bar{p}_i$$  \hspace{1cm} (24)

where the step length $\alpha_i$ is determined from minimization of the single-argument function

$$f(\alpha) = J(\bar{u}_i - \alpha \bar{p}_i) \rightarrow \min$$  \hspace{1cm} (25)

Direction vectors $\bar{p}_i$ are “conjugate” about matrix $A$ if they satisfy the following condition:

$$\bar{p}_i^T A \bar{p}_j = 0 \text{ for any numbers } i \neq j \text{ from the range } 1 \ldots n$$  \hspace{1cm} (26)

The methods of this family only differ in the way they construct the conjugate vector basis.

### 6.1 Analysis of the methods

All methods have been evaluated from the following standpoints: (i) rate of convergence; (ii) demand for computer memory; (iii) accuracy of the solution; (iv) requirements for smoothness of the target function and (v) the quality of conjugation of the descent vectors.

Numerical examples in the linear case have shown that the variable metric method has the highest rate of convergence (tending to quadratic). The theoretical convergence of the conjugate gradient method is violated by the instability of Gram – Schmidt orthogonalization procedure that it is based on. As a result, finite-step convergence after $N = n + n_1$ iterations was not achieved. This method is also very sensitive to the quality of unidirectional minimizations (Equation (25)) on each step. At the same time, variable metric and dual directions method do not require precise minimizations to maintain their rate of convergence; it suffices that the target function monotonically decreases from step to step.

The Powell method uses a very large number of unidirectional minimizations to construct the conjugate directions basis. This is unfeasible for the problem where the gradient can be computed almost as easily as the function itself. The rate of convergence is slow and the accuracy is very sensitive to the quality of the elementary minimizations.

The quality of vector conjugation is the highest in Powell method and is much worse in the gradient method. Our attempts to combine it with the Powell method, using conjugate directions constructed in one method in the other method, did not improve its performance.

As for the computer memory demand, conjugate gradient method is potentially the most economical one; in one of its modifications, only 3 vectors of dimension $N$ need to be stored. However, this advantage is largely lost because of the need to store the bulky transformation
matrix, $G$, for computation of the gradient. The variable metric method requires $\frac{1}{2} N(N + 1)$ memory cells for the metric matrix vs. the $N \times N$ array of conjugate directions (that all need to be stored) in the Powell method. However, in the Powell method, there is no need to form and store the $G$ matrix. The dual directions method has the highest demand for memory (two full $N \times N$ matrices need to be stored, plus the $G$ matrix). Over all, it was decided that the variable metric method is the most suitable method for the types of problems considered.

7 CONSTRUCTION OF STATICALLY ADMISSIBLE STRESS FIELDS

For plane stress problems, assuming that all loads and all supports are located at the nodes of the FE mesh, the static equations after discretization can be presented in coordinate form:

\[
\begin{align*}
\sum_{l=1}^{s} \left( \sigma^{l}_x I_{ml} + \tau^{l}_{xy} Y_{ml} \right) &= \sum_{r=1}^{s} P_{xr} \delta_{mr} \\
\sum_{l=1}^{n} \left( \tau^{l}_{xy} I_{ml} + \sigma^{l}_y Y_{ml} \right) &= \sum_{r=1}^{n} P_{yr} \delta_{mr}
\end{align*}
\]  

where $r$ = the number of loaded nodes including supports with unknown reactions, $\delta_{mr}$ is the Kronecker delta, and the coefficients $I_{ml}$ and $Y_{ml}$ are calculated as follows:

\[
\begin{align*}
I_{ml} &= \int_{A} \phi \frac{\partial \varphi_{m}}{\partial x} tdA \\
Y_{ml} &= \int_{A} \phi \frac{\partial \varphi_{m}}{\partial y} tdA
\end{align*}
\]  

The matrix of this system has the rank $2n$. The “fundamental” and “dependent” variables can be identified by reducing the system to the row-echelon form known from linear algebra, bringing it to the format of Equation (21). This decomposition need only be performed once and only the much lighter reverse pass is to be repeated on subsequent iterations.

If fracturing or cracking occurs, the number of equations in the system changes (some of the stresses reduce to zero) and some of the fundamental variables may become dependent. However, the number of these corrections will be small compared to the total order of the system, and it is beneficial to know that “almost all” of the system is already solved.

The corresponding algorithm has been developed for planar structures discretizable with a regular rectangular mesh. Detailed description of the algorithm is too lengthy to be placed here but the general idea is as follows. The system of static equations is decomposed into regular blocks and reduced to block row-echelon form, except for the bottom part that forms a resolvent subsystem of the order $2n_x \times (n + n_z + n_l)$ (here, $n_x$ is the number of grids in the $x$ direction, in which the nodes are numbered first). This is much less than the size of the original system (Equation (27)). In this subsystem, Gaussian elimination can be performed and fundamental and dependent variables identified. There will be exactly $2n_x$ dependent variables, including 3 support reactions required to assure global stability of the structure.

After the lower subsystem is solved, performing the reverse pass in the block form, matrices $A_f$ and $A_d$ from Equation (21) are constructed. The algorithm is completely universal and very computationally efficient. It reduces the solution of large Equation (27) to a series of subsystems of significantly smaller dimensions. Most of intermediate matrices need not be
stored in the computer memory but can be formed as necessary by finite formulae; others are represented as sets of compact cells, which in turn have strip diagonal structure.

8 CONSIDERATION OF CRACKS AND FRACTURING

Two separate conditions are considered in the model to represent local failure of concrete. When the first condition, $\Phi_1(\sigma_{ij}) \leq 0$ is violated, the concrete is deemed to crush and lose all load-bearing capacity; at this point, $\sigma_{ij} = 0$ is set. Violation of the second condition, $\Phi_2(\sigma_{ij}) \leq 0$ causes lineal rupture of the concrete. Neglecting aggregate interlock of the crack sides for simplicity, impose the conditions: $\sigma_n = 0$, $\tau_n = 0$ or, in the original coordinate axes,

$$\sigma_{ij} n_i n_j = 0; \quad \sigma_{ij} t_i t_j = 0$$

where $n_i$ = normal vector and $t_i$ = tangent vector to the surface of the crack.

The concept of a statically admissible stress field should now be amended to include the fulfillment of the strength criteria $\Phi_1(\sigma_{ij}) \leq 0$, $\Phi_2(\sigma_{ij}) \leq 0$. If, at any point of the structure, the strength condition has been violated, the following “operation” needs to be performed: (i) set $\sigma_{ij} = 0$ or impose Equation (29) at this point, depending on the type of violation; (ii) redistribute the residual stresses; (iii) exclude this point from further consideration (previously formed cracks are not deemed to close). This “operation” disrupts regularity of the process of solution; as a result of the “operation”, both the current point and the direction of descent may be lost and even the number of arguments of the target function may change. Solution of this highly non-trivial minimization problem requires special techniques.

Firstly, it is advisable to use a gradient-based method because the gradient of the target function can be computed inexpensively. Secondly, the descent should be performed only in the SASF domain. If the current point of descent on one of the iterations falls outside of this domain, the point should be projected back onto the domain and the descent resumed. This is the essence of the gradient projection method [7]. “Projection” consists of finding a point in the allowable domain that is closest to the current point. A theorem proven in [7] states that projecting does not alter the fact or the rate of convergence of the gradient method as linear.

Finding the exact projection is an optimization problem of the same complexity as the original problem, since the allowable domain is defined implicitly. In practice, when locating the exact projection is so difficult, it may suffice to perform an approximate projection by finding a point that belongs to the allowable domain and is uniquely defined by setting of the initial point, but is not necessarily the closest to it.

However, the cited theorem is only applicable for convex functions defined on convex domains, and with exact projections. For approximate projections, the proof of the theorem is not valid. When more advanced superlinear methods such as conjugate gradients or variable metric are considered that use information from previous iterations, the leaps associated with the “operation” of crack propagation may totally disrupt their convergence.

Fortunately, it is possible to prove that at least for some methods of minimization, “leaps” caused by cracking do not affect the fact or the rate of their convergence. The proof of this statement considers formation of one crack at a time. Suppose the minimization is performed using Equations (24) and (25). If $\alpha_k > \alpha_k^0$ (the length of the step that causes the cracking event
to occur), then, when passing through the point \( \alpha_k^0 \), both the function and its gradient sustain a leap and the properties of conjugate directions are lost.

In this situation, it is reasonable to stop at the point \( \alpha_k = \alpha_k^0 \), impose the constraint \( u_i = 0 \), re-establish the direction of descent and restart the process from the new initial point. However, this means that on the previous step, the exact minimum by \( \alpha_k \) has not been achieved, which may disrupt the convergence of the method.

Fortunately, there are methods such as variable metric and dual direction method that do not require precise elementary minimizations. These methods remain superlinearly convergent regardless of the step length, as long as the target function monotonically decreases from point to point on the descent path. Therefore, with proper minimization method, cracking in the structure can be accounted for without disruption of the process convergence.

9 THE ULTIMATE ANALYSIS

The purpose of the ultimate load capacity analysis of reinforced concrete structures is to determine their failure load value. Methods and programs of stepwise loading detected structural failure as the disruption of the algorithm of analysis, be it non-convergence or excessive displacements. Contrary to this, in the present paper the ultimate load is treated as the load under which the solution ceases to exist, regardless of the numerical method used.

The variational principle of Equation (17) states that if the solution of the problem exists, then it brings an absolute minimum to the functional of complementary energy. At the instant of failure, the conditions of existence of the minimum are violated. This may be caused by two different scenarios. Firstly, if the load is large, there may be no SASF equilibrating this load. As the load increases, the allowable domain in the hyperspace of stresses first contracts to a single point, and then disappears altogether. This situation is characteristic of the plastic bodies: for them, the failure load is the largest load under which a SASF still exists.

Brittle materials, however, may fail at a lesser load than the ultimate SASF load. Here, the second possible situation arises: the SASF domain still has some nonzero diameter but in this domain, the functional may not have an extremum. This requires considering deformation of the material in addition to its strength.

The conditions of static equilibrium after discretization will take the form of Equation (21). Under quasi-monotonic loading, the vector of external loads can be presented as

\[
\begin{align*}
\vec{P} &= \vec{P}_0 + \chi \cdot \vec{P}_1
\end{align*}
\]

where \( \vec{P}_0 \) and \( \vec{P}_1 \) are given (constant) vectors and \( \chi \) is the unknown scalar parameter of loading. Substituting and separating “fundamental” and “dependent” variables, obtain

\[
\begin{bmatrix}
\mathbf{A}_d & \mathbf{A}_f & -\vec{P}_1
\end{bmatrix}
\begin{bmatrix}
\vec{\sigma}_d \\
\vec{\sigma}_f \\
\chi
\end{bmatrix} = \vec{P}_0
\]

Since the load parameter \( \chi \) is unknown, it is logical to treat it as another “dependent” variable. Then, one more of the formerly dependent stress components should be moved to
fundamental variables. Once all $\sigma_f$ have been determined from the minimization process, the remaining $\sigma_d$ and $\chi$ can be obtained from Equation (31).

Therefore, the ultimate analysis of the given structure can be interpreted as a “maximin” problem or a double extremum problem: to maximize the (implicitly defined) function $\chi(\sigma)$ whose values are obtained by minimization of another function. The challenge here is that the values of the function are obtained by a very “expensive” calculation and the domain of definition of this function is not known a-priori. The target maximum will likely be attained at the edge of the domain. To address this problem, it is proposed to use the ideology of penalty functions method where the function’s domain of definition is expanded beyond the strength criterion (Equation (4)) but overloading of the structure results in disagreeable values of $\chi$, so the desired maximum cannot be achieved in the non-physical range.

10 CONCLUSIONS
- The proposed method is a mechanically consistent alternative to step-by-step FEA for simple paths of loading. It allows consideration of material nonlinear deformation, cracking, fracturing, and imperfect bond between concrete and steel.
- The proposed algorithm is essentially nonlinear, which conforms to the physical nature of the problem. At the same time, methods of linear algebra are used to effectively solve those of the equations (i.e. static equilibrium equations and boundary conditions) that are linear. This lays the basis for the new algorithms for crack formation and propagation and for the ultimate bearing capacity analysis.
- For the ultimate capacity analysis, a rational “maximin” procedure is proposed where previously, the algorithm failure was misinterpreted as the failure of the structure.

REFERENCES
CFD APPLICATIONS ON THE DESIGN AND ANALYSIS OF AERONAUTICAL PRODUCTS AT FADEA SA

GONZALO MOGNOL¹, DIEGO MAXIMILIANO CARRASCO²

¹ Flight Physics, Product Engineering, Fábrica Argentina de Aviones S.A
Av. Fuerza Aérea 5500 - X5010JMN - Córdoba - República Argentina
mognol@fadeasa.com.ar www.fadeasa.com.ar

² Flight Physics, Product Engineering, Fábrica Argentina de Aviones S.A
Av. Fuerza Aérea 5500 - X5010JMN - Córdoba - República Argentina
carrasco@fadeasa.com.ar www.fadeasa.com.ar

Key words: CFD, aircraft loads, control surfaces, high lift systems.

Abstract. Computational fluid dynamics (CFD) is a simulation tool, which uses powerful computers and applied mathematics to model fluid flow situations for the prediction of heat, mass and momentum transfer and optimal design in industrial processes. In this work results obtained in the calculation of aerodynamic loads with CFD on the wing of an aircraft (which underwent an engine change comparatively between the original plane, the modified one and calculations made with potential theory) are presented. In this sense, it was calculated the streamlines on both aircraft configurations, and then the aerodynamic behavior was checked when the flow goes around each type of nacelles. These numerical estimations reduce complexities and cost of using wind tunnel testing.

Another case study presented is the calculation of the aerodynamic loads on the flaps of this plane with the effects of the power plant included, comparing the levels of loads between the original aircraft and the re-engined one. The results of these calculations were the basis of a later analysis with FEM of the wing structure which is not presented herein. Moreover, a study of the diffusion of hot gases and their influence on the speed measurement probes in the aforementioned plane is presented.

As a final example, the design cycle of the control surfaces and flaps on a primary / basic training aircraft starting from the two-dimensional profiles up to the final results on the three-dimensional configurations is presented.

Last but not least, the importance of using commercial software and custom made software to reduce time and risks in the design process for an aeronautical product is highlighted.
1 INTRODUCTION

The main objective of this work is to present the use of Computational Fluid Dynamics (CFD) on the analysis and design of different aeronautical products at FAdeA SA. CFD is a computational tool which uses powerful computers and applied mathematics for the prediction of heat, mass and momentum transfer and optimal design in industrial processes.

In FAdeA SA several design processes and projects are carried out; the former require the use of high reliability for the prediction of aerodynamic loads, the characterization of aeronautical configurations and the prediction of heat and mass transfer. Commercial tools used in our company for this kind of simulations are ANSYS Fluent® together with other tools of pre and post processing belonging also to ANSYS.

The application examples that will be shown in this paper consist of the comparative calculation of loads over the wing and the high lift systems of an aircraft that underwent an engine change and the possible influence of the hot gases exiting the stubs of the new engine on the statics ports of the aircraft anemometric system. Last example that will be explained is the design cycle of the lateral control surfaces and high lift systems of a primary basic trainer using CFD and other custom made tools.

Finally, the importance of using CFD on the design process and analysis of aeronautical products in the industry will be emphasized from the viewpoint acquired in FAdeA.

2 COMPARATIVE CALCULATION OF THE AERODYNAMIC LOAD DISTRIBUTION OVER THE WING OF AN AIRCRAFT HAVING TWO DIFFERENT WING-MOUNTED ENGINES

The present case explains the application of CFD on the aerodynamic loads calculation for different maneuvers over the aircraft wing which is undergoing an engine change. This aircraft is a low wing aircraft with conventional fuselage having two turboprop engines mounted on the wing. The load case considered consists of a pull up with a rolling motion and is called Rolling Pull-Out maneuver.

The configuration analyzed consists of the main wing with the nacelles and the fuselage (Figure 1). Only this part of the aircraft is analyzed, because the influences of the horizontal and vertical tail are considered negligible on the load distribution over the wing in these maneuvers.

In the case of the calculations made for the asymmetric maneuvers (rolling Pull-Out) of this aircraft, the maneuver is subdivided into different components and the individual contribution of each one is calculated so that they can be combined following the principle of superposition and independence of effects. The components of the maneuver considered for the calculation of the aerodynamic loads are that of the symmetric flight, due to the deflections of the ailerons and the aerodynamic damping resulting from the rolling motion. Each one of these effects is simulated as a different maneuver with Fluent® obtaining the loads for each one.
The loads for the symmetric component are obtained simulating the aircraft in two different angles of attack. To discretize the computational domain, a tetrahedral mesh is used and then this mesh is converted to a polyhedral one because the numerical diffusion and the stability of the solution is better in this kind of meshes.

![Figure 1: Geometry of the aircraft with both nacelles for the aerodynamic analysis](image)

Due to the fact that the aim of this simulation is to calculate the aerodynamic loads and their distribution for different angles of attack, the mesh has a boundary layer mesh suitable for the simulation of the boundary layer through wall functions. This kind of model for the boundary layer is appropriate for the calculation of the aerodynamic loads when Reynolds number of the flow is high and the flow is attached to the wall. In Figure 2 the surface mesh of the aircraft with both nacelles is shown. The number of elements used in this mesh was about 3.27E6 with 1.05E7 nodes.

![Figure 2: Surface mesh of the aircraft with both nacelles](image)

The turbulence model selected for the simulations was the Spalart-Allmaras one, and the spatial discretization schemes used were of second order accurate for all the variables. To check the convergence of the solution, the continuity, velocity and turbulent species residuals and the force coefficients were used. The effort calculations are made for two different angles of attack (both in the range of linear effects) and then those values are combined linearly to obtain the loads from different angles of attack that the ones that were calculated. Calculations are carried out in this way for two reasons: on the one hand because one of the objectives of the original
work was to compare the load distribution obtained with CFD with the values obtained with the original Multhopp method used, and on the other hand because the latest method is linear, to avoid too much deviation from the linear theory the CFD calculation were made on the linear range of angles of attack too.

The following image shows load distribution per unit of dynamic pressure over aircraft main wing at an angle of attack of 5°. As it can be observed, the correlation between CFD simulations results and Multhopp theory is very good. However, some differences can be pointed out: mainly near the nacelles on the torsor moment. These differences occur because Multhopp theory does not take into account the presence of the nacelle on its calculations.

Figure 3: Comparison of the load distribution on the aircraft for $\alpha = 5^\circ$

Loads due to aileron deflections are calculated using a complete model of the aircraft with the ailerons deflected at a small angle. This angle of deflection is chosen to be small to avoid nonlinear effects in the simulations as detached flow. The mesh type used for the simulation was the polyhedral one with 8E6 elements and 1.8E7 nodes. Once the flow around the aircraft is solved for both nacelles, the load distribution is calculated and reduced per unit of dynamic pressure and per unit of deflection so that they are subsequently combined and it is possible to calculate the different maneuver loads. Figure 4 shows aircraft pressure contour with the ailerons deflected. As it can be observed pressure coefficient distribution is quite similar in both aircrafts showing only small differences nearby the nacelles.
For rolling damping effects calculations, a complete model of the aircraft is used. It is modeled carrying out a steady rolling maneuver using the MRF (Moving Reference Frame)\textsuperscript{[1]} model of Fluent\textsuperscript{®}. As well as in the case of the load calculation for aileron deflection and in the symmetric maneuver, the turbulence model used was the Spalart-Allmaras. Wall functions also were used to model the boundary layer and the spatial discretization schemes were of second order accurate for all the variables.

The following figure shows the pressure coefficient contours over the aircraft with both nacelles in the stationary rolling maneuver where more pronounced depression on the upgoing wing can be observed.

Once loads were available for each of the maneuver components and for the aircraft with both nacelles, they were linearly combined to obtain the total load of each particular maneuver that implies a load factor (or angle of attack) an aileron deflection and a rolling velocity. Subsequently, results of these loads were used for the verification of the wing structure on the design maneuvers with the new engine.
3 COMPARISON OF THE LOADS OF THE HIGH LIFT SYSTEM OF AN AIRCRAFT WITH DIFFERENT ENGINES

This analysis case consists of the assessment of the influence of the change in the nacelle geometry and the propeller on the high lift system loads of a twin turboprop aircraft. The trigger that led to this study was the consequence of the following fact: when changing an engine, flow conditions behind the propeller changes and as the flaps are partially submersed in the slipstream of the propeller, it was necessary to reassess the loads to know whether there is a level increase or not.

The aircraft geometry evaluated was the wing+fuselage+nacelles configuration with a propeller model consisting of an actuator disk. Flap deflection was set to the maximum possible for the aircraft and the power setting of the engines to that necessary of the maneuver taking into account. The mesh type used for the simulations was the polyhedral one with 3.5E6 elements and 9.2E6 nodes.

The propeller modelling is done by means of an actuator disk, that is to say a disk which has the same diameter as that of the propeller and in which a discontinuity is set to the static pressure, to obtain the desired thrust. To solve the flow, the SST-κω turbulence model was used because it is suitable for solving flow with high adverse pressure gradients and separated flow, as can be possible with flaps at high angles of deflection.

With the data of the velocities, flight altitudes, angles of attack and thrust required, the simulations of the airflow on the plane with both engines are made. From these simulations the loads on the structure of the flaps are obtained and compared to each other and with the original load calculations, to determine whether or not it is necessary to perform a verification of it.

Figure 6 shows pressure coefficient contours on the aircraft with both nacelles at α = 0° and take off power while the Figure 7 shows aircraft wall shear stress streamlines. Considering both configurations it can be noticed that there is a difference in the pressure distribution and flow behavior in the area of the inner flap, which led to further analysis of its structure to check the levels of internal stresses.
In this study case it is analyzed the possible influence of the engine change on a twin turboprop aircraft on the lateral static ports of the aircraft anemometric system. Possible influences that were evaluated were: first, influence of the nacelle geometry change in the static pressure on the static ports; second, influence of the propeller diameter change in the flow conditions behind it and third, influence of hot gases exiting the engine stubs on the temperature nearby static ports. To establish whether those influences exist or not, simulation of the aircraft with the operation of the engine were carried out comparing the pressure and temperatures on the static ports of the aircraft with both nacelles.

To accomplish the simulations, several flight situations are considered taking into account different altitudes, velocities and dissimilar power settings. Then, some of the cases are selected since they are considered the most critical from the point of view that the influences on the static ports could be the maximum and the simulations with the Fluent® software are carried out. The cases that resulted to be the most critical are three level flight conditions with maximum power and with the maximum yaw angle attainable by the aircraft.

Since the simulation has to be with the aircraft having a yaw angle, the geometry of the aircraft that must be used is that of the complete aircraft. On the other hand, as it was stated previously, the simulations require the consideration of the engine operation, therefore in the nacelle geometry the air inlet and the exhaust gas stubs are modeled so as to establish the appropriate boundary conditions for the inlet air and hot gases exit. Besides of that, the propeller operation was simulated through an actuator disk which introduces a discontinuity in the static pressure on the disk representing the propeller.

The computational mesh used for this simulation was a polyhedral one and the number of elements used was 4.36E6 with 1.26E6 nodes. To solve the flow a steady solver is used with a coupled algorithm for velocity-pressure coupling. The spatial discretization schemes were of second order accurate for all the variables. The flow is considered compressible and the energy...
equation was solved to account for the hot gasses diffusion. The flow exiting the engine stubs is considered air with the mass flow and temperature appropriate to the power setting of the engine and the flight condition considered. Finally, the turbulent model used was the Spalart-Allmaras with wall functions to model the boundary layer.

Once simulations were carried out, the temperature and pressure on the lateral static ports for both aircrafts were compared obtaining the differences shown in the Table 1. As it can be observed the differences in pressure and temperature in the ports on both aircraft are negligible in the cases analyzed.

<table>
<thead>
<tr>
<th>CASE</th>
<th>$\Delta P$ (%)</th>
<th>$\Delta T$ ($K$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.021</td>
<td>-0.25</td>
</tr>
<tr>
<td>2</td>
<td>-0.075</td>
<td>1.17</td>
</tr>
<tr>
<td>3</td>
<td>-0.073</td>
<td>1.32</td>
</tr>
</tbody>
</table>

Table 1: Comparison of temperature and pressure between both aircrafts

Figure 8 shows temperature contours around the aircraft and the streamlines exiting the engine exhaust stubs colored by temperature comparing the aircraft with both engines. As it can be seen in that image, the temperature in the fuselage surface is almost the same as the ambient temperature for the flight condition considered and hot gases high temperature exiting engines stubs is diffused in the air a short distance behind the stubs and it does not reach static ports.
In conclusion, we can say that in the simulations carried out it was found that neither the nacelle geometry nor the exhaust gases temperature of the new nacelle induce substantial change in the static pressure and temperature of the static ports at the fuselage sides.

5 DESIGN CYCLE OF THE LATERAL CONTROL SURFACES AND HIGH LIFT SYSTEMS OF A PRIMARY/BASIC TRAINER USING FLUENT®

Another application of CFD in FAdeA is in the design cycle of high lift systems and the lateral control surfaces, in this particular case in a primary/basic trainer.

High lift systems design implies defining several parameters of two-dimensional geometry since they affect the most the slot efficiency to control the flow through it and over the flaps itself.

Those geometrical parameters are the flap size, the position and the shape of the slot lip, the entry shape (with or without skirt, with or without curvature), the flap nose shape, the flap position regarding the slot lip and the hinge axis position.

To determine the quality of the flow in the slot it is necessary to evaluate whether there areas of the flap with separated flow, if a stream acceleration occurs in the upper surface of the flap, if the direction of flow at the exit of the slot "wets" the upper surface of the flap or whether the flow from the lower to the upper surface of the profile is sufficient to energize the boundary layer of
the flap. In order to obtain the desired performance of the flaps for the range of angles of attack and deflections, several geometries were evaluated with Fluent® software. Each geometry was assessed using a bidimensional model of the flap and it was obtained the force coefficients for several angles of attack.

Once defined the domain around the profile, a mesh that is refined as we approach the wall is created. To capture velocity gradients in the boundary layer a prismatic mesh is created around the main profile and the flap, using 24 layers with a growth rate of 1.2 and a first layer thickness of 2.3x10⁻⁵C. Once the domain is meshed, it is possible to load the turbulence model, boundary conditions, solvers, etc. As turbulence model, SST-κω was used which provides accurate predictions in separate flow problems with adverse pressure gradients.

Finally, a wing was analyzed with this two-dimensional flap design. This three-dimensional geometry of the flap was generated by extruding the two-dimensional geometry under analysis taking into account the chord variation along the wingspan flap. Besides, the geometry of the slot was generated with the same premises. The configuration tested was that of WING (with 25°-deflected flap) + FUSELAGE.

In the light of the results obtained through three-dimensional simulations, the required lift in any flight situation desired was analyzed including landing, when a high level of lift is needed.
These results determine the minimum speed of the aircraft with flaps deflected, landing speed and maximum landing weight. When the flap design goal is to maintain a desired minimum aircraft speed with flaps, to increase aircraft weight is necessary to implement another command mechanism or a different slot shape.

In the case of the design of the lateral control surface for a primary/basic trainer, once the airfoils of the wings, the position of the hinge line and the surface needed for the aileron are selected it is necessary to design the type of aileron that best fits the requirements for the aircraft.

The types of ailerons that this aircraft can use are several and each one has its advantages and disadvantages. In the case of this aircraft it was decided to use fies type ailerons which has the advantage of facilitating mass and aerodynamic balance and also reduces the adverse yaw of the aircraft. Considering this type of aileron, the geometry of the slot, the nose of the movable surface and the hinge line position is varied until the best design is obtained.

To evaluate the lift, drag, moment and hinge moment coefficient of the designs for different angles of attack and deflections the Fluent® software is used. The geometry evaluated is the two-dimensional one with different deflections and angles of attack. To do so, a computational domain and a mesh are constructed as it can be seen in Figure 12. The simulations are implemented considering a Reynolds number of 4E6, the solver used is steady and the algorithm of pressure-velocity coupling is coupled, the turbulence model used is the SST-κω and the spatial discretization schemes used is of second order accurate for all the variables.

![Figure 12: Computational domain and mesh used for the evaluation of the two-dimensional airfoil with aileron deflections.](image)

Once the two-dimensional results are available they are transferred to the three-dimensional wing through a program that uses the Multhopp equations and allows estimate in a quick manner the rolling velocity, the adverse yaw moment and the stick forces for a given dynamic pressure, load factor and aileron deflection.

The parameters that are changed to meet the requirements of rolling velocity and yaw moment, keeping the force on the stick at acceptable levels, i.e. they are not too low or too high, are the maximum deflection of the left and right ailerons, the relationship between the two deflections and the transmission ratio of the ailerons with the stick.
Once all the requirements are met, the final design of the lateral control system is finished.

6 CONCLUSIONS

At present the CFD tools are been widely used in the aeronautical industry assisting on the design and analysis of aeronautical products all over the world. At FAdeA this tool is relatively new but we visualize a path of rapid and steady growth in optimizing our designs through its use.

One of the most promising advantages identified in the use of CFD assisting in the design processes is the capability of evaluating several designs quickly and with high reliability greatly reducing long hours spent on modeling and wind tunnel analysis.

At FAdeA we have a large number of wind tunnel tests result that allowed us to achieve promising results adjusting the tool and encourage us to extend and generalize the use CFD in the aerodynamic design process.

7 REFERENCES

COMPARISON BETWEEN THE DRAG POLAR CURVES USING ANALYTICAL AND NUMERICAL METHODS IN CONTROLLED RADIO AIRPLANE

ANDERSON P. ALMEIDA\textsuperscript{1}, LUIZ M. PEREIRA\textsuperscript{2}, CAIO CESAR G. SANTOS\textsuperscript{3}, LEONARDO S. MACHADO\textsuperscript{4}

\textsuperscript{1}Universidade Federal do Vale do S\~{a}o Francisco (UNIVASF) Laboratory of Computational Fluid Dynamics, Juazeiro, BA, Brazil anderson2999@gmail.com

\textsuperscript{2}Universidade Federal do Vale do S\~{a}o Francisco (UNIVASF) Laboratory of Computational Fluid Dynamics, Juazeiro, BA, Brazil luj.mariano@univasf.edu.br

\textsuperscript{3}Universidade Federal do Vale do S\~{a}o Francisco (UNIVASF) Laboratory of Computational Fluid Dynamics, Juazeiro, BA, Brazil caio_cgs@hotmail.com

\textsuperscript{4}Universidade Federal do Vale do S\~{a}o Francisco (UNIVASF) Laboratory of Computational Fluid Dynamics, Juazeiro, BA, Brazil leosampaio_fsa@hotmail.com

Key words: drag polar curve, CFD, analytical method, numerical method, radio controlled airplane, maximum takeoff weight.

Abstract. This paper aims to compare the drag polar curves between numerical method and analytical method of a controlled radio airplane and identify its results at the maximum takeoff weight. It is presented a theory about Fluid Mechanics and Aerodynamics, in addition to the boundary conditions and the software used in the CFD simulations. It was observed by the results that both methods resulted in similar drag polar curves, but the curve obtained in the CFD simulations resulted values much closer in practice for small and high angles of attack. It was obtained a relative error between the $CD$ and $CL$ using both methods below 10\% for most of the speed range, as well as a relative error of the 3.5\% and 7.9\% in the prediction of the maximum takeoff weight using the numerical method and analytical method, in comparison to the maximum takeoff weight obtained in flight tests, respectively. The analytical method can be used in the preliminary design while the numerical method can be used to predict with more accuracy the $CD$ and $CL$ values, and finally resulting in a better prediction of the airplane performance in conditions of the takeoff, landing, and straight flight.

1 INTRODUCTION

Every body immersed in a flow of a fluid is subjected to the forces exerted by the fluid. It is no different than an aircraft, which will be always subject to airflow, thus being exposed to aerodynamic forces. Among the aerodynamic forces experienced by an aircraft, the lift and the
drag are the most important. The first as a force directed perpendicularly to the flow and responsible for maintaining the aircraft in the air. The later directed in parallel and opposite direction to the movement of the aircraft.

The aerodynamic forces are represented by dimensionless coefficients: lift ($C_L$) and drag ($C_D$), having the information about the ability of a given geometry to generate their forces.

One way to visualize the relationship between the lift and the drag on an aircraft is provided in the drag polar curve, which consists of a graph in polar coordinates relating the drag coefficients and lift coefficients of all aircraft components, plus all information about the aerodynamic forces experienced by the complete aircraft.

The drag polar is of great importance for the design of an aircraft, it provides all the information necessary for the performance analysis of the same, so the polar calculations must be done in a way that best represents the aerodynamic coefficients.

The polar calculation can be done using numerical and analytical methods, the last one being suggested by different authors. These methods differ themselves in the way that the lift and the drag coefficients for each aircraft component is calculated. The numerical method can be performed using CFD, English short for Computational Fluid Dynamics. This tool is based on the computational solution of the equations of mass, momentum and energy conservation that govern the flow [2]. Differently of the analytical method where the coefficients are calculated for each aircraft component separately, the CFD techniques provides a direct result of the complete aircraft aerodynamic coefficients.

The CFD tool offers a reliable and low cost alternative for solving complex problems. Therefore, it has been very important in the current aeronautical projects, in particular the calculation of the drag polar.

2 TURBULENCE MODELS

In CFD simulations, one of the main problems is to model the turbulent flow once in this kind of problem the flow is irregular and changes randomly with time and space. In this way, the turbulence models modify the equations and just consider the average of turbulence effects. In the presented case, the software used was STAR-CCM+®, and some turbulence models was analyzed and tested until be used in final simulations.

2.1 Spalart-Allmaras Turbulence Model

The S-A models solve a single transport equation that determines the turbulent viscosity. The model was developed for the aerospace industry, and its advantage is that it can be readily implemented in an unstructured CFD solver. It is mainly recommended for attached boundary layers and flows with mild separation, such as flow past a wing, cases in that the best results are noted [7]. As the generated simulation was of lesser complexity and due to computational constraints, this was the turbulence model used in this case, providing reliable results, similar to was expected. The transport equation solved in this model is presented by [3] as:

$$u_x \frac{\partial \tilde{v}}{\partial x} + u_y \frac{\partial \tilde{v}}{\partial y} = c_{b1} \tilde{v} + \frac{1}{\sigma_x} \left[ \nabla \times (\nu + \tilde{v}) \nabla \tilde{v} + c_{b2} (\nabla \tilde{v} \times \tilde{v}) \right] - c_{w1} f_{w1} \frac{\tilde{v}}{d}$$

(1)
2.2 K-Epsilon Turbulence Model

The K-ε is a model based in two equations that solves the transport equations for the turbulent kinetic energy (K) and its dissipation rate (ε). Many variations of K-ε model have been used by decades, and it is the most used for industrial applications. Since its inception, have been many attempts to improve it. At first, the K-ε turbulence model was applied with wall functions, fact that was changed later when was modified to use Low-Reynolds number and Two-layer for resolving the viscous sublayer [7].

2.3 K-Omega Turbulence Model

The K-ω model is a two-equation model and an alternative to the K-ε turbulence model. In this case, the transport equations are solved for the turbulent kinetic energy (K), and a quantity ω, that is the specific dissipation rate, or in other words, the dissipation rate per unit of the turbulent kinetic energy.

One advantage of this model over the K-ε is its improved performance for boundary layers under adverse pressure gradients. Furthermore, the K-ω may be applied throughout the boundary layer, even the viscous-dominated region, without changes.

The main disadvantage in this model is that boundary layer computations are sensitive to the values of ω in the free stream that means great sensitivity to inlet boundary conditions for internal flows [7].

3 PROBLEM SPECIFICATION

The main dimensions of the airplane are visualized at the annex. It was used two methodologies for plotting the drag polar curves, the numerical method, using CFD analysis, and analytical method recommended by [5]. The main stream air velocity was defined as the range 14-27 (m/s) and the air properties are taken at 600 m, which is the location where the airplane should fly.

According these methodologies to calculate C_D and C_L coefficients the velocity must be used. The airplane was drawn on SolidWorks® and imported to STAR-CCM+® to obtain the numerical simulations. There are three types of meshes in STAR-CCM+: tetrahedral, polyhedral and trimmer. The trimmer mesh allows a better control and concentration of cells in specific regions of the domain, as an example, the region called wake. Various volumetric controls were used to support the mesh refinement at the regions where the pressure and velocities gradients are stronger, according Figure 1.

![Figure 1: Refinement zones near the airplane](image)
At the refinement zones, the cells sizes were chosen between 6 and 10 mm. To support the solution of the boundary layer equations, it was used the resource prism layer mesher, which allows a better refinement near the walls. Figure 2 shows this resource.

The domain was sized to minimize the wall effects at the flow field near the airplane. It was chosen a cubic domain (34.5 x 34.5 x 34.5 m) seen in the Figure 3.

The Standard Spalart-Allmaras Turbulence, SA, combined with the All $y^+$ Wall Treatment was selected to calculate the aerodynamic characteristics of the airplane. In this case, values of $y^+ \leq 1$ includes a viscous sublayer good resolved. In this work were used the equations 1 and 3.
suggested by [4] and [3], respectively, to calculate the height of the first cell near the wall and the total thickness of the prisms near the wall.

\[ y = \frac{y^+ \cdot \nu}{u^*} \]  

Where \( u^* \) was obtained by:

\[ u^* = 0.05 \cdot U \]  

\[ \delta = 0.37 \cdot D \cdot \left( \frac{1}{R_e} \right)^{\frac{1}{5}} \]  

Where \( y \) is the height of the first mesh cell near the wall, \( y^+ \) is the nondimensional distance to the wall, \( u^* \) is the friction velocity and \( \nu \) is the kinematic viscosity. In this work, the stop criteria used calculates the difference between the maximum and the minimum values of \( C_D \) and \( C_L \), collected during the last 30 samples. The simulation is completed when the difference results in the value is \( 1 \times 10^{-3} \).

4 RESULTS AND DISCUSSION

4.1 Velocity

In the Figures 4 and 5, the velocity magnitude is shown in the symmetry plane of the airplane for two different angles of attack (\( \alpha \)). It was observed a great wake behind the fuselage and over the wing in these two cases.

![Velocity magnitude](image)

**Figure 4:** Velocity magnitude at the symmetry plane to \( \alpha = 13.5^\circ \)
4.2 Mesh Refinement

According the wall treatment chosen it was adopted a reference value $y^+ = 1$, and was obtained the height value of the first prismatic cell near to the wall and the total thickness according equations 1 and 3. $y^+ \leq 1$, to the wall treatment conditions used in this work, indicates that the mesh is sufficiently refined to solve the boundary layer. The Figure 6 illustrates $y^+$ distribution over the airplane.

4.3 Lift ($C_L$) and drag ($C_D$) coefficients

After the simulations and the analytical calculations, $C_L$ and $C_D$ plots for the entire airplane were found, as shown in the Figures 7 and 8.
Was calculated the relative error between $C_L$ and $C_D$ values obtained with the two methods, for each velocity, according the equation:
relative error = \left| \frac{φ_{analytical} - φ_{numerical}}{φ_{numerical}} \right| \cdot 100\%

Where \( φ \) is the \( C_D \) and \( C_L \) values for each method and velocity airplane. The values obtained are shown in the Figures 9 and 10.

**Figure 9:** Relative error of \( C_D \) for different airplane velocities

**Figure 10:** Relative error of \( C_L \) for different airplane velocities

### 4.4 Drag polar curve

The \( C_D \) and \( C_L \) values, obtained in the CFD simulations, was calculated through the equations:
Where $C_L$ is the lift coefficient, $C_D$ is the drag coefficient, $q_\infty$ is the dynamic pressure and $S$ is the planform wing area.

According each $C_L$ and $C_D$ calculated was plotted the curve shown in Figure 11, that represents the drag polar curve for the complete airplane.

Figure 11: Drag polar curve for the complete airplane

As can observed through Figure 11, the curves have a similar behavior, however the curve obtained through the numerical method, result in larger values of $C_D$ for low $C_L$ values. After to analyze the results was shown that when the $C_L$ becomes lower, for velocities above 19 m/s, the parasite drag grows faster.

After the drag polar results were made the takeoff performance analyze, that includes, the maximum takeoff weight. To verify what drag polar curves were closest of the maximum weight obtained in fly tests were used the equations below [1]:

\[
S_g = \frac{1}{2gK_a} \ln \left( 1 + \frac{K_A}{K_T} V_{L0}^2 \right) + NV_{L0}
\]  
(8)

\[
K_T = \frac{T}{W - \mu r}
\]  
(9)
\[ K_A = -\frac{\rho_\infty}{2(W/S)} \left[ C_{D,0} + \Delta C_{D,0} + \left( k_1 + \frac{G_s}{\pi eAR} \right) C_L^2 - \mu_r C_L \right] \]  

Where \( S_g \) is the runway length for takeoff, \( g \) is the gravity, \( K_A \) and \( K_t \) are constant values, \( V_{L0} \) is the velocity of takeoff, \( N \) is the time necessary to the roll of the airplane until the complete takeoff, \( T \) is the engine thrust, \( W \) is the airplane weight, \( \mu_r \) is the static friction coefficient, \( \rho_\infty \) is the specific density of the air, \( C_{D,0} \) and \( \Delta C_{D,0} \) are the parasite drag, \( k_1 \) is a constant value, \( G_s \) is the ground effect coefficient, \( e \) is the wingspan efficiency factor and \( AR \) is the aspect ratio.

Through those equations were calculated the maximum takeoff weight, shown in the Table 1.

<table>
<thead>
<tr>
<th>Method</th>
<th>Total Weight (N)</th>
<th>Relative error* (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analytical</td>
<td>106,6</td>
<td>7,9</td>
</tr>
<tr>
<td>Numerical</td>
<td>95,4</td>
<td>3,5</td>
</tr>
<tr>
<td>Flight tests</td>
<td>98,9</td>
<td>-</td>
</tr>
</tbody>
</table>

* Relative error in comparison to the weight obtained in flight tests.

The Table 1 shows that was obtained a lower relative error in the maximum takeoff weight estimation through the numerical method in comparison with the analytical method.

5 CONCLUSIONS

This paper has simulated flow over an airplane. The simulation was conducted in STAR-CCM+® with the Standard Spalart-Allmaras turbulence model. The drag polar curves obtained have a same behavior, but the drag polar curve obtained through the CFD simulation has a behavior closest according the drag polar curve obtained in some references. The detachment of the boundary layer and the growth of the wake region are not as well modeled in an analytical method than numerical method. There is an increase of the parasite drag that is not modeled as well using the analytical method. Through the maximum takeoff weight values is observed that the drag polar curve using the numerical method resulted in a minor error on the analytical analysis when compared with the actual value obtained in flight tests.

REFERENCES

ANNEX – Airplane dimensions
PARALLEL COMPUTATIONS FOR MANY FLOATING OBJECTS TRANSPORTED BY TSUNAMI FLOWS

K. Aoki¹, S. Ushijima², H. Itada³ and D. Toriu⁴

¹ Department of Civil and Earth Resources Engineering, Graduate School of Engineering, Kyoto University
Kyoto, 615-8540, Japan
aoki.kazuma.34z@st.kyoto-u.ac.jp

² Academic Center for Computing and Media Studies (ACCMS), Prof., Kyoto University
Kyoto, 606-8501, Japan
ushijima@media.kyoto-u.ac.jp

³ Department of Civil and Earth Resources Engineering, Graduate School of Engineering, Kyoto University
Kyoto, 615-8540, Japan
itada.hiroshi.34n@st.kyoto-u.ac.jp

⁴ Department of Civil and Earth Resources Engineering, Graduate School of Engineering, Kyoto University
Kyoto, 615-8540, Japan
toriu.daisuke.57x@st.kyoto-u.ac.jp

Key words: Parallel computation, Floating objects, Collision, Tsunami

Abstract. In the Great East Japan Earthquake in 2011, the huge tsunami flows caused massive damage to the coastal areas in the wide range of Tohoku regions. In particular, it is pointed out that a large number of tsunami debris and drifts increase the devastating damage. Thus, it is important to understand the behavior of various many drifts for disaster mitigation. For that purpose, a numerical method is taken as a useful method. In this study, a computational method MICS [1], which is applicable to incompressible multiphase field consisting of gas, liquid and solid phases, is applied to the wave-induced flows including many floating objects through static structures and complicated grand shapes. As a result of numerical experiments, it can be seen that the MICS is effective to predict the free-surface flows with the floating objects among structures and on non-uniform ground surface.
1 INTRODUCTION

Great East Japan Earthquake in 2011 caused extensive damages in the wide range of Tohoku regions. Especially, the floating objects transported by tsunami flows caused serious damage for the structures. Thus, it is an important subject to numerically predict the behaviors of the transported objects as well as the free-surface flows.

As the computational methods for floating objects by tsunami, Yoneyama et al.[4] and Kawasaki et al.[5] investigated the computational method that predicts a few floating objects. However, the computational method to deal with a large number of floating objects as found in actual tsunami disasters has not been proposed so far. This is because the difficulties to treat the collisions of objects, mechanical interactions between different phases and computational load to calculate such plenty of objects.

In this study, we calculated the free-surface flow including a large number of arbitrarily shaped objects by the multiphase model MICS [1] and attempted fast calculation with parallelization by MPI. The main features of the present computational method are shown as follows:

- Since the fluid-solid interactions are adequately taken into account, no empirical constants are needed.

- Mainly principal governing equations are described with conservative forms, which are discretized with finite volume method (FVM) on a collocated grid system.

- C-ISMAC method enables us to utilize the high-order TVD schemes and to use larger time increment \( \Delta t \) than explicit methods.

- To discuss the applicability the numerical method, we solved large-scale problems predicting the behavior of more than 200 objects. In order to solve the problems efficiently, the computations were parallelized by flat MPI on the basis of a domain decomposition method.

2 NUMERICAL PROCEDURES

In this study, we employed MICS proposed by Ushijima et al.[1] as the prediction method to solve the multiphase field of the gas, liquid and solid phases. In addition, an arbitrarily-shaped objects is represented by tetrahedron elements, whose collisions are treated by the contacts of the multiple “collision detection spheres” [8] placed near the surface of the objects on the basis of DEM [7].

2.1 MICS

Firstly, we calculate the velocities and pressures of the gas-liquid field and the fluid forces which affect objects. Next, we calculate the motions of objects and then accommodate the effects of objects to gas-liquid field. The principle governing equations of
gas-liquid field are mass conservation law, incompressibility condition and equation of motion. They are shown as follows:

\[
\frac{\partial \rho_f}{\partial t} + \frac{\partial}{\partial x_j}(\rho_f u_j) = 0 \quad (1)
\]

\[
\frac{\partial u_j}{\partial x_j} = 0 \quad (2)
\]

\[
\frac{\partial u_i}{\partial t} + \frac{\partial}{\partial x_j}(u_i u_j) = f_i - \frac{1}{\rho_f} \frac{\partial p}{\partial x_i} + \frac{1}{\rho_f} \frac{\partial}{\partial x_i} \left[ \frac{\partial}{\partial x_j} (\mu u_i) + \frac{\partial}{\partial x_i} (\mu u_j) \right] \quad (3)
\]

where \( t \) is time, \( x_i \) is component of the orthogonal coordinate system and \( f_i \) is external acceleration of \( x_i \) component. \( u_i \) is the mass average velocity of gas and liquid phases. \( p \), \( \rho_f \) and \( \mu \) are volume average pressure, density and coefficient of viscosity. These equations are discretized on the collocated grid system based on a finite volume method, and the convection term is treated by means of fifth-order TVD scheme [3]. In order to solve these principle equations, we employed SMAC method. In the prediction stage, we use implicit method (C-ISMAC method) [2], and in the pressure calculation stage, we use C-HSMAC method [6] to fulfill the incompressibility condition sufficiently.

In this study, we correct the velocities of gas and liquid by means of the ratio of objects to fluids in the cells and calculate the collisions of objects with each other. In order to calculate motions of the arbitrarily-shaped objects, the objects in the fluids are treated as rigid bodies and represented by multiple tetrahedron elements.

The tetrahedron elements are used for calculation of volume fractions of objects in each cell. Collision detection spheres (CDS) [8] are placed around the inside of the surfaces. The CDS are used to calculate the contact forces, which occurs when the objects collide with each other or any wall surfaces.

The ratio of objects to fluids are calculated by means of tetrahedron sub-cell method [9]. In this method, we divide a fluid-computation cell into multiple sub-cells and evaluate the volume ratio occupied by an object from the numbers of the sub-cells included the cell. The contact forces are estimated by DEM [7] with CDS.

The fluid forces which affect the objects are calculated by volume integral for pressure and viscosity term in right hand side of Eq.(3). Here, a volume of \( m \)-th tetrahedron of block \( k \) is \( T_{km} \), and a part or all of \( T_{km} \) included in the cell \( C \) is defined as \( \Delta T_{Ckm} \). \( i \) component \( F_{Ckm,i} \) of fluid force vector which affect \( \Delta T_{Ckm} \) is obtained by the following equation, where \( \rho_{bk} \) is the density of the block \( k \).

\[
F_{Ckm,i} = \rho_{bk} \Delta T_{Ckm} \left[ -\frac{1}{\rho} \frac{\partial p}{\partial x_i} + \frac{1}{\rho} \frac{\partial}{\partial x_j} \left\{ \frac{\partial}{\partial x_j} (\mu u_i) + \frac{\partial}{\partial x_i} (\mu u_j) \right\} \right] \quad (4)
\]
2.2 Calculation of Collision with Objects

In this study, we treat calculation subjects as aggregates of rigid bodies on basis of DEM [7] and solve the equation of motion of each element along time evolution considering the collisions with each other. The rigid bodies are called contact detection spheres (CDS) in the following.

For the sake of simplicity, we employ linear springs and consider only dynamic friction force in tangential direction, not considering static friction. The tangential and normal direction force are shown as follows:

\[
f_n = -K_n d_n n + D_n v
\]

\[
f_t = K_t v_{rt} \Delta t + D_t v_{rt}
\]

where \(K_t\) and \(D_t\) are the spring coefficient and the linear viscous damping coefficient of dash pot in tangential direction. \(K_n\) and \(D_n\) are ones in normal direction.

In order to treat the collisions of objects, we place contact detection spheres inside the object surfaces. The center of the sphere is located to coincide with the center of gravity of the tetrahedron, and the radius is decided to satisfy that the volume of the sphere should be equivalent to the volume of the tetrahedron.

Firstly, the relative velocity vector \(v_r\) of sphere \(j\) of block \(l\) to sphere \(i\) of block \(k\) is shown as follows:

\[
v_r = v_l - v_k + \omega_l \times r_{l,ij} - \omega_k \times r_{k,ij}
\]

Here, \(v_l\) and \(v_k\) are the velocity vectors of each block, and \(\omega_l\) and \(\omega_k\) are the angular velocity vectors of it. \(r_{l,ij}\) and \(r_{k,ij}\) are the vectors from the center of gravity of each block to the point of tangency of sphere \(i\), \(j\). The velocity vectors of tangential and normal directions to contact surface are obtained from \(v_r\), and the contact forces are obtained from Eq.(5) and (6).

The resultant and moment of the contact force to block \(k\) written as \(F_k\) and \(N_k\) are shown as follows:

\[
F_k = \sum_{i \in \Psi_k} \sum_{i \in \Omega_i} f_{ij}
\]

\[
N_k = \sum_{i \in \Psi_k} \sum_{i \in \Omega_i} r_{k,ij} \times f_{ij}
\]

where \(f_{ij}\) is the vector which occurs by the collision between sphere \(i\) and \(j\) and then affect sphere \(i\). \(\Psi_k\) is the set of the spheres included in block \(k\). \(\Omega_i\) is the set of ones which contact sphere \(i\).
2.3 Parallelization

In order to solve this large scale problem efficiently, the computational was parallelized by flat MPI (Message Passing Interface) on the basis of a domain decomposition method. In this method, the domain is divided into some sub-domains. The calculations in each sub-domains are conducted as processes which have independent memories. If you parallelize fluid calculations, you need velocities and pressures of neighbor sub-domains.

In this study, the domain is divided in $x_1$, $x_2$ and $x_3$ direction. Positive direction of $x_3$ is vertically upward. The domain is divided by $l$ in $x_1$ direction, $m$ in $x_2$ direction and $n$ in $x_3$ direction. It is parallelization to $l \times m \times n$, and $1 \times 1 \times 1$ is the sequential computation.

![Figure 1: Parallelization by domain decomposition method](image_url)
3 APPLICATIONS OF COMPUTATIONAL METHOD

3.1 Floating Objects Transported by Tsunami Flows

In order to confirm that the present computational method is able to predict the transportation of many floating objects by tsunami flows, the numerical experiments are conducted.

The calculation area is shown in Fig.2, which is $l_1 = 2.0$ [m], $l_2 = 1.0$ [m] and $l_3 = 2.0 \times 10^{-1}$ [m]. The ground surface is set as shown in Fig.3. The ground surface is treated as a static object. The height of the ground surface $h(x_1, x_2)$ was artificially set up as follows:

$$h(x_1, x_2) = \max \{0, f(x_1, x_2)\} + \max \{0, g(x_1, x_2)\} \quad (10)$$

$$f(x_1, x_2) = -\left\{ (x_1 - l_1)^2 + (x_2 - l_2)^2 \right\} + \frac{4}{5} l_3 \quad (11)$$

$$g(x_1, x_2) = -\left\{ (x_1 - 0.7l_1)^2 + x_2^2 \right\} + \frac{4}{5} l_3 \quad (12)$$

The working fluids are air and water, whose density are $1.0 \times 10^3$ [kg/m$^3$], and viscosity are $1.0 \times 10^{-5}$ [m$^2$/s], $1.0 \times 10^{-6}$ [m$^2$/s], respectively. Objects information of vehicle and structure models is shown in Tab.1. The spring constants of tangential and normal direction $K_t$ and $K_n$, which is used to calculate contact force, are determined as $K_t = K_n = 1.0 \times 10^4$ [N/m]. The viscous damping constants of dashpot $D_t$ and $D_n$ are determined as $D_t = D_n = 4.1$ [N $\cdot$ s/m] regarding coefficient of restitution as 0.3 [10].

Figure 2: Calculation area

Figure 3: Setting of ground surface
Figure 4: Configuration of objects on horizontal plane

Figure 5: Setting of objects

Table 1: Object information

<table>
<thead>
<tr>
<th>Object</th>
<th>Vehicle model</th>
<th>Structure model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node number</td>
<td>172</td>
<td>354</td>
</tr>
<tr>
<td>Number of Tetrahedron elms.</td>
<td>399</td>
<td>1320</td>
</tr>
<tr>
<td>Number of CDS</td>
<td>276</td>
<td>492</td>
</tr>
<tr>
<td>Object density [kg/m³]</td>
<td>$3.57 \times 10^4$</td>
<td>$2.56 \times 10^4$</td>
</tr>
</tbody>
</table>

The initial conditions are given as follows: Water mass which height is $0.4l_3$ is set on $x_1 \leq 0.3l_3$. The numbers of vehicle and structure models are 240 and 40 respectively. The configuration of objects on a horizontal plane is shown in Fig.4, in which coarse and fine meshes show process and cell configurations. The diameter of columns is $C = 9.0 \times 10^{-2}$ [m] and one column is composed of 4 structure models.

The boundary conditions are given as follows: A non-slip boundary condition is imposed on all walls. On the vertical walls $\partial p/\partial x_1 = 0$ and $\partial p/\partial x_2 = 0$, while $p$ is set by considering gravity on the horizontal walls.

The number of computational cells is $400 \times 200 \times 40$, and the number of parallel processes is $20 \times 10 \times 2$. The number of computational cells per process is $20 \times 20 \times 20$. The time increment $\Delta t$ is $5.0 \times 10^{-4}$ [s].

3.2 Predicted Results

The predicted results from $t = 0$ [s] to $t = 3.8$ [s] are shown in Fig.6. It can be seen that the vehicle models are reasonably transported by tsunami flows and collide with moving
and static objects. Especially, the transportation of the floating objects is interfered by the static column structures as shown in Fig.6 (c). On the other hand, the transportation of the floating objects along ground surface far from the static column structures is not affected by the structures. To prevent the movement of the floating objects, it is proposed that some additional columns structures should be set along the ground surface.

Figure 6: Calculated drifts and free-surface flows with static columns (top views)

Fig.7 shows the time history of the maximum number of objects $N_p$ included in one
It can be seen that \( N_p \) increases around at \( t = 1.0 \) [s], and that the number of objects in one process is biased. Applying the dynamic load-balancing parallel computation method proposed by Maruyama et al. [11], it is expected that the method could enabled us to decrease computation load efficiently.

The efficiency of the parallel computation was examined by changing the core numbers. Fig.8 shows the speed-up ratios on the basis of 8 cores in the Cray-XE 6 in Kyoto University (AMD Opteron 2.5GHz, 32 cores / node, 64GB memory / node). The maximum core number is 400. As shown in Fig.8, the tendency of the increasing speed-up ratio is near the linear line. Thus, the present flat MPI parallelization enables us to decrease the elapse time satisfactorily.

\begin{figure}[h]
\begin{center}
\includegraphics[width=0.4\textwidth]{figure7.png}
\caption{Maximum number of objects per process}
\end{center}
\end{figure}

\begin{figure}[h]
\begin{center}
\includegraphics[width=0.4\textwidth]{figure8.png}
\caption{Speed-up ratios of flat MPI}
\end{center}
\end{figure}

### 3.3 Influence of Static Structures

In order to confirm how the transportation of the floating objects is affected whether static columns are set or not, the numerical experiment without columns has been conducted. The computational conditions except the columns are same as 3.1. The predicted results from \( t = 0 \) [s] to \( t = 3.8 \) [s] are shown in Fig.9. It is successfully predicted that the tsunami flows make floating objects move downwards according to the complicated-shaped ground surfaces. Thus, it was shown that the effects of static column structures on the transport processes, in which the movements of floating objects are obviously different from those in Fig.6, are reasonably predicted by the present method. This effect is clearly shown in the following result; the floating objects are interfered by static columns at \( t = 1.6 \) [s] in Fig.6, while Fig.9 shows that they flow down with the tsunami flows.

From the above results, the present computational method enables us to estimate how the static columns prevent the transportation of many floating objects by tsunami flows. Thus the computational method is utilized to mitigate the damage by floating debris.
4 CONCLUSIONS

In this paper, numerical experiments have been conducted with MICS [1], which is a computational method for incompressible multiphase fields. In particular, a parallel computational method is employed in order to deal with a large number of floating objects with sufficient resolutions by fluid-computation cells.

As a result of the computation of numerical experiments, it has been shown that
MICS [1] enables us to predict the behavior of the floating objects, which collide with other moving and static objects, transported by tsunami flows on complicated-shaped ground surfaces. It was also confirmed that the static objects disturb the transportation of floating objects by tsunami flows.

REFERENCES


PRACTICAL CRITERIA FOR NUMERICAL MODELING OF WATER QUALITY IN COASTAL ZONES

ANGEL N. MENÉNDEZ¹,², NICOLÁS D. BADANO¹,², EMILIO A. LECERTÚA, AND MARIANO RE¹,²

¹ INA, Hydraulics Laboratory
Autop. Ezeiza-Cañuelas, Tramo J.Newbery km 1,62 (1804) Prov. Buenos Aires, Argentina
angel.menendez@speedy.com.ar http://laboratorios.fi.uba.ar/lmm

² Univ. Buenos Aires, Eng. Dept., LaMM
Av. Las Heras 2214, (1127) Ciudad Buenos Aires, Argentina

Key words: Water Quality, Numerical Errors, Numerical Bias, Practical Criteria.

Abstract. In order to establish limited water use zones in the coastal area, and/or define the values of the design parameters of the disposal system, numerical modeling must be used. This includes a hydrodynamic model, to simulate the circulation in the coastal zone, and a pollutant transport model, to determine the spatial and temporal concentration distributions. The numerical aspects of both types of models are already well established. But selection of the appropriate ones for the particular problem, and proper implementation and validation of the selected models require some expertise. In the present paper, based on a combination of conceptual principles and experience, a series of practical criteria are stated in order to choose, implement and validate those models. They are illustrated through their application for two case studies of particular relevance in Argentina: water quality in the coastal zone of the city of Buenos Aires, on the Plata River, including the project of a new set of two outfalls, and the design of an outfall for the city of Ushuaia on the Beagle Channel, in Patagonia.

1 INTRODUCTION

Water quality in coastal zones of urban settlements is usually impacted by domestic discharges. In order to mitigate or remediate this impact, an extended enough sewage system must be implemented, including water treatment plants. In many cases, the discharges from the plants cannot be localized at the coast itself if water quality needs to be preserved throughout the coastline; hence, subaquatic outfalls must be built.

Numerical modelling is presently the main tool to establish limited water use zones due to pollution in the coastal area, and/or define the values of the design parameters for the sewage disposal system [1,2,3,4,5,6,7]. This includes a hydrodynamic model, to simulate the circulation in the coastal zone, and a pollutant transport model, to determine the spatial and temporal concentration distributions.

The numerical approaches for both types of models are already well established. Hydrodynamic models are usually based on finite differences [8], finite elements [9], or finite volume [10] formulations. Pollutant transport models can be of the eulerian [11] or lagrangean [12] types.

However, selection of the appropriate models for the particular problem, and proper
implementation and validation of the selected models require some expertise. In the present paper, based on a combination of conceptual principles and experience, a series of practical criteria (best practices) are stated in order to choose, implement and validate those models.

They are illustrated through their application to two case studies of particular relevance in Argentina: water quality in the coastal zone of the city of Buenos Aires, and the design of an outfall for the city of Ushuaia, in Patagonia.

2 NON NUMERICAL ASPECTS

When dealing with environmental problems, the uncertainty in the input data and in the theoretical models for the transformation processes (chemical, biochemical and biological reactions) impose limits to the accuracy of the results obtained from numerical modelling. Moreover, the goals of the coastal water quality studies, associated with the delimitation of zones apt for certain uses, do not require a high level of accuracy. This does not mean that numerical errors are not relevant, but that their reduction must be referred to the former sources of error and the required overall accuracy.

Water quality problems in the coastal area usually aim to: (i) establish limited water use zones (i.e., zones not prone to be used for human consumption, recreation or biota preservation) under existing coastal use conditions; (ii) define the values of the design parameters for the sewage disposal system (outfalls extension, level of treatment for the water treatment plants, water quality control of the tributaries) when planning urban expansion or dimensioning remediation measures, in order to achieve previously established water uses.

Now, the criteria to determine if a zone is apt for a definite water use are usually of a statistical nature. Specifically, it is common to require that the concentrations of a definite set of water quality parameters remain simultaneously below a respective set of threshold values for most of the time (usually 95% of the time). This condition can be checked at specific points from measurements only if a continuous and sufficiently long record is available for each point. In order to determine a limited use zone from measurements, a dense net of monitoring stations should be implemented [13], which constitutes a high-cost strategy. Obviously, numerical modelling is the most appropriate tool to perform this checking, once properly calibrated.

The ideal situation to calibrate the water quality model (see next section for more details) is to have available relatively long concentration records at many points, to which the numerical predictions can be compared. Unfortunately, this type of data does not usually exist. Even if it were, it must be taken into account that the numerical model requires as input data the pollutant loads from the main sources, information difficult to get with precision due to economic and practical reasons.

Regarding the transformation processes models, they are stated as decaying/growing reactions of the first order type, or the Monod type, with attenuation factors which decrease fast when Dissolved Oxygen (DO) content gets too low [14]. These empirical formulations are limited from the theoretical point of view, and the values of their associated parameters could usually be chosen within wide uncertainty ranges.

Additional non numerical sources of error exist, though their influence is, in principle, small. In the first place, in coastal problems the shallow water approximation (hydrostatic pressure) holds, as water depths are small relative to the horizontal scales of variation for the
water currents. In the second place, though presently three-dimensional (3D) models are ready to be used for engineering purposes, application of the two-dimensional (2D) approximation (vertically integrated models) constitutes the proper choice for the far field when vertical mixing is effective enough to smooth out any stratification.

3 NUMERICAL ASPECTS

3.1 Numerical errors

The main source of numerical errors is the spatial and temporal discretization of the differential equations. However, the selection of the model domain and the associated open boundary conditions, also constitute significant sources of error.

As already mentioned, controlling numerical errors in the model results means that their magnitude should be small relative to the ones arising from input data errors, model errors, and acceptable levels of errors for water management decision making. But, in particular, one aspect should be carefully addressed concerning numerical errors: avoid biases, i.e., concentrate the effort in controlling systematic more than random errors. This is especially relevant for the pollutant transport model, as discussed below.

3.2 Model domain

Obviously, the model domain should be large enough to include the whole problem-zone. But, frequently, the crucial issue, from the point of view of numerical errors, is the location of the open boundaries, which are only mathematical borders where – sometimes relatively rough – approximate boundary conditions must be imposed (which could then introduce biases), especially for the hydrodynamic model.

The selection of the location of the open boundaries should be based on physical grounds, i.e., choosing those borders where the soundest approximations can be made for the boundary conditions. In any case, buffer zones adjacent to the open boundaries should be considered, with extensions large enough to smooth out the influence of the details of those approximate boundary conditions on the problem-zone.

3.3 Spatial grid

A good representation of the hydrodynamics is essential to avoid biases in the pollutant transport model, even in spite of using relatively poor transformation processes models. This is so because advection tends to be a controlling process in water quality problems, determining the path of pollution, as it acts faster than the other ones.

From the point of view of the hydrodynamic model, the spatial grid must be chosen so as to properly capture the velocity gradients. A rule of thumb could be to have no less than 10 nodes to capture ranges of velocity differences between adjacent local maxima and minima; 20 nodes would be optimum; larger number of nodes would be useless, with the penalty of increasing roundoff errors. As in coastal zones velocity gradients are usually linked to bathymetric gradients, the spatial grid can be built based on a proper representation of the bathymetric surface (i.e., the above criterion can be applied to bottom elevation data).

When relatively extended buffer zones between open boundaries and the problem-zone exist, the accuracy criterion for the spatial grid could be relaxed in those buffer zones – where
precision is not an issue – without penalizing the accuracy within the problem-zone. Eventually a nested model approach could be used [15], in order to generate, from the parent model, the boundary conditions for a child model encompassing only the problem-zone.

A proper representation of horizontal mixing (through vertical axes turbulent eddies, and/or the combination of vertical differential advection and horizontal axes turbulent eddies – so called ‘dispersion’ – in vertically integrated models) is a key issue to properly simulate pollutant transport without introducing bias.

Mixing is mathematically represented through Fick’s law, leading to a diffusion term in the pollutant balance differential equation. If an eulerian approach is used for the pollutant transport model, numerical diffusion in the horizontal plane (arising from the discretization of advective terms) should be controlled in order to maintain it low relative to the physical diffusion, especially in the direction normal to the velocity, in order to avoid a bias towards higher diffusion. This is achieved by reducing the spatial step, usually much below the one selected for the hydrodynamics model close to pollutant discharge points, in order to properly solve the pollutant concentration gradients.

The eulerian approach could be too costly for some applications. In this case, the lagrangean approach should be adopted, which does not impose further restrictions on the spatial step (relative to the one selected for the hydrodynamic model), unless higher precision for the concentration distribution – obtained through post processing of the model results – is required (in which case, alternatively, a denser post processing grid could be built instead of densifying the model grid [16]).

3.4 Time step

The selection of the time step, $\Delta t$, in order to get accuracy, depends on the spatial grid. In principle, for the hydrodynamic model, the time step should be of the order of the time needed for the tidal wave to travel a distance of the order of the smallest spatial step, i.e.,

$$Cr \equiv \frac{c \Delta t}{l} \sim 1$$

where $Cr$ is the Courant Number, $c$ the tidal wave celerity (including the flow velocity), and $l$ the scale for the spatial step. Numerical stability requirements, associated to explicit time discretization methods, may lead to specific constrains on the time step (Courant condition). When the tidal wave propagation is not an issue (domain extension very short in comparison to wavelength), implicit methods are more appropriate, avoiding numerical stability constrains, and allowing $Cr$ values much larger than 1.

In the case of a pollutant transport model of the eulerian type, the time step should be of the order of the time needed for the flow to travel, and the diffusion to spread over a distance of the order of the spatial step, i.e.,

$$Cr' \equiv \frac{U \Delta t}{l} \sim 1$$

$$\Delta t \sim \frac{l^2}{D}$$

where $Cr'$ is a second Courant Number, $U$ the flow velocity, and $D$ the diffusion coefficient. As with the hydrodynamic model, using an explicit method would lead to specific


constrains on the time step. For a pollutant transport model of the lagrangean type, the time step between particle injections should be controlled so as to generate a sufficiently dense cloud of particles. As a rule of thumb, at least 10 particles per cell should be present in each averaging cell within the impacted zone.

3.5 Model validation and calibration

In order to test and adjust the model, a calibration scenario must be chosen. This is essentially a time window for which all model input data are known (or can be inferred), and the system response (flow velocities and pollutant concentrations) at certain points was measured.

A model is considered as validated if it qualitatively and quantitatively captures the main features of the system response for the calibration scenario, previous to being calibrated. Were not the case, the missing or misrepresented mechanisms should be identified, and introduced or corrected, until it performs appropriately.

Once model validation is achieved, calibration must be undertaken. This means tuning the model parameters values, within their uncertainty ranges, in order to achieve the best possible agreement with observations for the calibration scenario. Goodness of fit indicators should be used as measure of that agreement. As a preliminary help for the calibration process, sensitivity runs should be performed in order to identify the key parameters to be tuned. Though the calibration process usually proceeds through level-headed trial and error, techniques for automatic calibration are available [17].

Sometimes, a verification scenario is chosen, which corresponds to a different time window with known model input data and system response. Model verification consists in comparing the observations for the verification scenario with the predictions of the already calibrated model, and showing that it performs appropriately, i.e., with values for the goodness of fit indicators similar to the ones obtained for the calibration scenario.

4 APPLICATIONS

4.1 Buenos Aires coastal zone

Buenos Aires is the Capital city of Argentina. The Metropolitan Region of Buenos Aires (MRBA) is adjacent to the coast of the Plata River, a very wide estuary (around 50 km at the city location) with a sea-like hydrodynamics, with tides and waves, though with freshwater in its Inner Zone (where the MRBA stands) due to the big discharges from its two main tributaries, the Paraná and Uruguay Rivers, the former through its two main branches, ‘Paraná de las Palmas’ and ‘Paraná Guazú’ (Figure 1). The Plata River waters by the MRBA, called the ‘Southern Coastal Fringe’ (SCF), are polluted due to the urban discharges. The SCF is a very long reach, of the order of 80 km, and extends laterally for a few kilometers.

Studies were conducted in order to establish limited water use zones for the present situation, and to predict their changes after the implementation of the Remediation Master Plan of the Water Company (AySA) for the MRBA.
In order to properly locate the open boundaries, the model domain included the whole Inner Plata River, with an extension of about 175 km (Figure 1). On the one hand, the upstream open boundaries were the mouths of the three big tributaries, where the corresponding daily time series of discharge can be inferred. On the other hand, the downstream open boundary locates at a cross section where hourly time series of water level are available at both coasts (a smart interpolation is used in between), and lies far enough from the problem-zone. At this section the estuary expands abruptly, and the hydrodynamic regime changes (lateral water currents become significant, and a saline wedge develops).

In the Inner Plata River, flow depths are relatively low (average of about 5 m, with maxima of the order of 10 m) relative to its horizontal dimensions and the tidal wavelength (about double the model length). Hence, the shallow water approximation holds. Low depths, absence of salt water, relatively low suspended sediment concentrations (of the order of 200 mg/l), and persistent action of winds combine to avoid conditions for stratification, so the two dimensional approach can be safely used.

Proprietary software Mike 21, from DHI Water & Environment, was used, which applies a finite volume method. It includes module ECOLAB, to deal with pollutant transport.

The spatial scale of variation of the natural topography (without taking into account the dredged navigation channels, considered as small disturbances relative to the study scale) is of the order of 5 km. An irregular, unstructured spatial grid with triangular elements was built. In the buffer zones (out of the SCF), spatial steps between 1.0 and 1.7 km were used. Within the SCF the spatial step was varied from about 200 m down to less than 100 m.

Due to the large quantity and wide spatial distribution of pollutant sources (Figure 2), an eulerian approach was selected for pollutant transport. No refinement of the spatial grid was performed due to computer restraints (at the time of the study), allowing for some numerical diffusion.
The time step is conditioned by the tidal wave celerity, of the order of 7 m/s (through Lagrange formula), leading to about 14 seconds for a spatial step of 100 m. However, a lower value was taken (3.5 seconds) to account for some smaller size elements around discharge points.

The validation of the hydrodynamic model was performed based on four different scenarios: (i) 20 days of a synthetic pure astronomical tide / comparison with Tide Table at Buenos Aires; (ii) one year (1997) simulation / comparison with frequency distribution curve for water level in Buenos Aires; (iii) 40-day period (2004) simulation / comparison with continuous velocity records at two stations (Figure 3); (iv) 5½ hours (2004) simulation / comparison with trajectory of a buoy under wind action. There were no uncertain parameters for which results were significantly sensitive, so no calibration was undertaken.

The available data to undertake the calibration of the pollutant transport model arose from 10 surveys performed at different seasons and years during the 1990 decade. Each survey consisted in the collection of instantaneous samples at 11 cross sections (see Figure 2), with 3 to 6 points per section, for a large set of parameters. From this database, 8 parameters (DO, BOD, fecal coliforms, Ammonium, Nitrate, phenolic substances, Chrome, and Lead) were selected, and their mean value and corresponding total range of variation were obtained. Additionally, similar surveys for the 24 discharge points from the MRBA to the Plata River (Figure 2), including an outfall (d190), provided mean values and ranges of variation for the pollutant loads.
The hydrodynamic and pollutant transport models were run for a time window of 20 days (March 2004) during which the discharge from the Paraná de las Palmas River (the one directly influencing the coastal zone) was close to its mean value (3,900 m$^3$/s), using the mean pollutant loads obtained from the surveys. The statistical distributions of concentration for the locations of all monitoring points were obtained from the model results, and compared with the variation ranges from observations. The calibration consisted in adjusting the reaction coefficients so as to achieve the best possible agreement between them (Figure 4).

This application illustrates a frequent practical situation: availability of data strongly conditions the calibration scenario. In other words, the practitioner has to adapt to existing data in order to validate and calibrate his models.

A verification of the pollutant transport model was undertaken a posteriori, through comparison of monitoring data obtained during surveys performed during 2008, and the corresponding results of the calibrated model [18]. The measured concentrations of the instantaneous samples mostly lied within the higher probability range indicated by the model.

The model was applied to determine present limited use zones, as illustrated in Figure 5. It was later used to determine their changes under the influence of the Remediation Master Plan [18].
4.2 Ushuaia coastal zone

Ushuaia is the world southernmost city. It lies on the northern bank of the Beagle Channel, in Patagonia, which connects the Atlantic and Pacific Oceans (Figure 6a). The coastal hydrodynamics is driven by the tide and the winds. Presently, sewage is discharged at a point located at the West of the peninsula, close to the coast. As part of a Master Plan for Water Provision and Sewage Disposal, two outfalls are projected; the main one emerging from the southern coast of the peninsula, and a secondary one at the eastern zone of the bay. Studies were undertaken to determine the impact of these discharges on the water quality of the Beagle Channel.

![Figure 6: Problem-zone and model domains](image)

A model domain encompassing the whole width of the Beagle Channel, of the order of 10 km, was selected to properly simulate the hydrodynamics (Figure 6b). The distance between the West and East open boundaries was about 40 km. Note that a smaller South open boundary also existed at Murray Channel. As the tidal wave travels from East to West, the tidal time series on the East boundary was specified as boundary condition. The West border was treated as a non-reflecting boundary, using Blumberg-Kantha radiation scheme, which performs a weighted average between the internally calculated wave (at the neighboring internal node) and the incident wave (at the Eastern boundary, with a time lag), using as weight a relaxation time to be specified. Salinity and temperature values are also imposed at the Eastern border.

The maximum flow depths are of the order of 200 m, much lower than the tidal wavelength and the model dimensions. Hence, the shallow water approximation can be adopted. According to previous studies, no significant stratification naturally occurs at the problem-zone around the peninsula. A two-dimensional approach was then used, ignoring any possible stratification produced by the discharge from the outfall at this first stage of the studies.

The simulations were performed applying the public domain numerical system MOHID, developed by MARETEC. It is based on a finite volume method on a regular mesh, allowing automatic nesting.

A nested approach was selected. Figure 6b shows the parent and child model domains. Variations of natural topography have a spatial scale of the order of 1 km. Hence, for the parent model a spatial step of 120 m was selected. It was reduced down to 30 m for the child
Due to the fact that only point pollutant sources were considered, a lagrangian approach was applied for pollutant transport simulation, using Allen-Sullivan formulation. This was performed on the child model.

As the model extension is small relative to the tidal wavelength, a Courant Number much higher than 1 can be chosen. The maximum Courant Number was set to 10, from which time steps of 12 and 4 seconds arose for the parent and child hydrodynamic models, respectively.

The time step for particle injection was selected as 1 second, leading to a particle density per cell varying between approximately 20 and 80.

Two calibration scenarios were used for the hydrodynamic model: (i) 15 days of a synthetic pure astronomical tide / comparison with Tide Table at Ushuaia harbor; (ii) 3 days of a mean annual tide / comparison with velocity hodographs obtained from measurements processing at 19 stations. The following values were finally selected for the calibration parameters: relaxation time = 250 seconds; time lag = 130 seconds; effective roughness height = 10 mm. Figure 7 illustrates the agreement between observed and simulated velocity hodographs; note that the velocities are relatively small, which amplifies the errors for both observations and simulation.

![Figure 7: Comparison of measured and observed velocity hodographs](image)

No calibration was performed for the pollutant transport model, as no data was available. Three water quality parameters were considered: fecal coliforms (representative of bacterial pollution), BOD (organic matter pollution), and TSS (physical pollution). First order reactions were assumed, with values for the decay parameters fixed on reasonability grounds.

Figure 8 shows the fecal coliform plumes associated to the main outfall for two instants of time at opposing tidal phases, while Figure 9 presents the resulting limited water use zones with the outfall project, considering only the three selected water quality parameters.

5 CONCLUSIONS

Conceptual and practical non numerical and considerations have been formulated for numerical simulation of water quality in the coastal zone.

Controlling numerical errors in the model results means that their magnitude should be small relative to the ones arising from input data errors, model errors, and acceptable levels of
errors for water management decision making.

Though the main source of numerical errors is the spatial and temporal discretization of the differential equations, the selection of the model domain and the associated open boundary conditions also constitute significant sources of error.

Biases should be avoided concerning numerical errors. Good representations of the hydrodynamics and of horizontal mixing are key issues to properly simulate pollutant transport without introducing bias.

![Figure 8: Simulated fecal coliforms plumes from main outfall for two instants of time](image)

![Figure 9: Limited use zones](image)

REFERENCES


Numerical modeling of heterogeneous media for the compressional wave using high-order finite-differences

ALEJANDRO GUILLEN, FRANCISCO H. CABRERA, JAIRO A. TORRES AND JESUS E. DURAN

PANGEA
Research Group in Geophysics and Geology
Universidad de Pamplona
Ciudad Universitaria, Km 1 Via Bucaramanga, Pamplona, Colombia
guillen-18@hotmail.com, fcabrera61@yahoo.com.br, jatorresp@gmail.com &
masentendimiento@hotmail.com

Key words: Finite-difference, Acoustic wave equation, Seismic modeling, High-order approximation

Abstract. A numerical solution to the wave equation in heterogeneous acoustic media is presented in this research article. The differential equation for the compressional field was written in terms of spatial derivatives and derivative from of spatial operator time. Initially, it presents a general method to achieve the approximations for approximating the operator spatial differentials of second order. This approximation is calculated using the values of the field at a point P of the computational grid and its adjacent points according to the degree of approximation used. It was subsequently implemented in a computational algorithm where the initial conditions of the field of P were set; a seismic source was implemented; and the corresponding numeric scheme was developed. The temporal part of the derivatives were solved using a second order approximation through centered operators. In the spatial derivatives, different orders of approximation were used starting with a second, fourth, sixth, eighth and tenth order, respectively. The conditions of stability and numerical dispersion were treated to prevent error growth exponentially and generation of numerical artifacts. The graph of numerical dispersion for the different values for the passage of spatial and temporal mesh were presented. Finally the numerical solutions using different orders of approximation on simple geological models compared.

1 INTRODUCTION

The finite difference method, traditionally used for solving the equation of acoustic an elastic wave [1,2,3,4,5,6,7]. The method has been successfully used to simulate the propagation of the seismic wave in inhomogeneous media, generally anisotropic and highly
complex geological media [8, 9, 10, 11]. Usually the time derivates is approximated by a centered second order finite difference operator. The Laplacian operator is approximated by a difference scheme of second or fourth order. The theory shows that satisfied the condition of stability, numerical dispersion (ie the difference between the phase velocity calculated by scheme and the actual speed) decrease when a greater number of points in the calculation of the Laplacian operator is taken.

If the finite difference scheme order (2nd, fourth, sixth, etc) is increased, the numerical dispersion decreases. In this context two parameters are fundamental: the step size and order spatial mesh finite difference scheme used. This paper shows that better results are obtained when the order of the scheme is higher when the step is fixed spatial mesh. In some details the steps further shown the coefficients of the finite difference scheme for any order, based on the method of undetermined coefficients, which is not another thing that direct application of Theorem Taylor [12]. The number steps of time for numerical simulation is determined by the step size of spatial mesh. For a relatively coarse mesh, it takes a relatively long time step, which means less processing time when a finer mesh is taken. However, in the first case the number may be great dispersion, it is to say is dispersion is a numerical artifact which obscures the simulation results, in particular synthetic seismograms obtained. For purposes beyond the modeling of seismic wave propagation, such as seismic migration [13] is important balance between computation time and accuracy required in the results. The finite difference scheme of high order are a good alternative for this purpose, since its implementation is simple flexible.

2 WAVE EQUATION AND THE METHOD OF FINITE DIFFERENCES

Wave equation is presented for two-dimensional case considering an acoustic isotropic media. This equation is written in terms of second derivates of order with respect to time and space, as shown [14]

$$\frac{\partial^2 P}{\partial t^2} = c^2(x, z)\nabla^2 P,$$

(1)

where $\nabla^2 = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial z^2}\right)$, also $P$, is the compressional wave field and is a function of spatial coordinates and time, therefore $P = P(x, z; t)$, $c(x, z)$ is the velocity of the medium. Important applications of the wave equation shown above ranges from Exploration Seismology, Seismic modeling to, Migration and Seismic Inversion.

In this paper the wave equation presented in (1) is a second order differential equation in the variable time. This equation is solved using the finite difference method. This operator is used to approximate the spatial derivatives and the derivative with respect to time. Using the 2nd order in time and higher-order finite differences, the forward propagation can be calculated as like:
\[ \begin{align*}
P_{i,j}^{n+1} &= 2P_{i,j}^n - P_{i,j}^{n-1} + \Delta t^2 c_{i,j,k}^2 \left\{ \left( \nabla^2 \right)^p \right\} P_{i,j}^n, \\
\end{align*} \tag{2} \]

where \( P_{i,j}^n \) is defined

\[ P_{i,j}^n = P(i\Delta x, j\Delta z; n\Delta t), \tag{3} \]

here the index \( n \) is associated with the time steps, and the indices \( i \) and \( j \) associated with the coordinates \( x \) and \( z \) respectively, which take values \( i = \{1, 2, \ldots N_x\}, j = \{1, 2, \ldots N_z\}, \) and \( n = \{1, 2, \ldots N_t\}. \) Also \( N_x \) y \( N_z \) is the number of spatial samples and \( N_t \) the number of samples of the time, with their mesh sizes \( \Delta x, \Delta z \) and temporal sampling interval \( \Delta t \)

The Laplacian with \( p \)-order of accuray is used in terms of the coefficients \( A_S \), which is shown below [16]

\[ \left\{ \left( \nabla^2 \right)^p \right\} P_{i,j}^n = \sum_{s=1}^{2p+1} A_s \left\{ \frac{1}{\Delta x^2} \left( P_{i-\left(p-s+1\right),j}^n \right) + \frac{1}{\Delta z^2} \left( P_{i,j-\left(p-s+1\right)}^n \right) \right\}. \tag{4} \]

The solution method for calculating the coefficients, is known as the method of undetermined coefficients shown in the next section. The stability condition for isotropic modeling is a follow [15]:

\[ \Delta t < \frac{\Delta d}{v_{\text{max}} \sqrt{\mu}}, \quad \mu = \sum_{m=-p/2}^{m=p/2} \left( \|w_x\| + \|w_z\| \right), \tag{5} \]

where \( \Delta d = \min(\Delta x, \Delta y, \Delta z) \) and \( v_{\text{max}} \) is the maximum velocity in the medium and both with the condition of stability allow to establish takes into account the condition of numerical dispersion for the scheme

\[ \Delta d = \frac{v_{\text{min}}}{(S) F_N} \tag{6} \]

where \( v_{\text{min}} \) is the minimum speed at which the wave propagates in the medium, \( F_N \) is the Nyquist frequency and \( S \) is the number of points needed for the propagation is nondispersive.
3 METHOD OF UNDETERMINED COEFFICIENTS

Briefly explain the method of undetermined coefficients for calculating the weights that accompany the discretized fields.

For \( p = 1 \) in the sum of equation (4), takes only the derivative with respect to the spatial coordinate \( x \) an approximation of the derivative at a second order, obtaining the following modelo (7)

\[
\frac{\partial^2 P}{\partial x^2} = \frac{1}{\Delta x^2} \left( A_1 P_{i-1,j}^n + A_2 P_{i,j}^n + A_3 P_{i+1,j}^n \right).
\]

(7)

For \( p = 2 \) an approximation of the derivative at a fourth order, resulting setting is

\[
\frac{\partial^2 P}{\partial x^2} = \frac{1}{\Delta x^2} \left( A_1 P_{i-2,j}^n + A_2 P_{i-1,j}^n + A_3 P_{i,j}^n + A_4 P_{i+1,j}^n + A_5 P_{i+2,j}^n \right).
\]

(8)

For \( p = 3 \) an approximation of the derivative at a sixth order, the resulting expression is

\[
\frac{\partial^2 P}{\partial x^2} = \frac{1}{\Delta x^2} \left( A_1 P_{i-3,j}^n + A_2 P_{i-2,j}^n + A_3 P_{i-1,j}^n + A_4 P_{i,j}^n + A_5 P_{i+1,j}^n + A_6 P_{i+2,j}^n + A_7 P_{i+3,j}^n \right).
\]

(9)

Taking into account the Taylor Series, let \( f(x) \) a function that has at least \( (n + 1) \) as a result an resulting in an interval \([a, b]\), where \( n \geq 0 \). and \( x_0, x \in [a, b] \)

\[
f(x) = P_n(x) + \underbrace{R_{n+1}(x)}_{\text{error}},
\]

(10)

where

\[
P_n(x) = \sum_{k=0}^{n} \frac{f_k(x_0)(x-x_0)^k}{k!},
\]

(11)

developing equation (11) for the term \( n \)

\[
P_n(x) = f(x_0) + \frac{f'(x_0)(x-x_0)}{1!} + \frac{f''(x_0)(x-x_0)^2}{2!} + \ldots + \frac{f^n(x_0)(x-x_0)^n}{n!}
\]

(12)

\[
R_{n+1}(x) = \frac{(x-x_0)^{n+1}}{(n+1)!} f^{(n+1)}(\xi),
\]

(13)

such that \( \xi \) is between \( x \) and \( x_0 \), and \( \xi \) is the error of running the inaccuracy of the Taylor series. Now, we can write the second derivative of \( f(x) \) function with respect to the spatial variable using different notations
\[ f''(x) = \frac{\partial^2 f}{\partial x^2} = D_n^{(2)} f(x). \]  

(14)

So it can be shown the second derivative in terms of the coefficients (weights) \( A_1, A_2 \) and \( A_3 \)

\[ D_n^{(2)} f(x) = A_1 f(x - h) + A_2 f(x) + A_3 f(x + h), \]  

(15)

using the Taylor series

\[ f(x - h) = f(x) - \frac{h}{1!} f'(x) + \frac{h^2}{2!} f''(x) - \frac{h^3}{3!} f'''(x) + \frac{h^4}{4!} f''''(x)(\xi_-), \]  

(16)

\[ f(x + h) = f(x) + \frac{h}{1!} f'(x) + \frac{h^2}{2!} f''(x) + \frac{h^3}{3!} f'''(x) + \frac{h^4}{4!} f''''(x)(\xi_+), \]  

(17)

substituting the equations found in (16) and (17) into (15) we obtain

\[ D_h^{(2)} f(x) = A_1 \left[ f(x) - h f'(x) + \frac{h^2}{2!} f''(x) - \frac{h^3}{3!} f'''(x) + \frac{h^4}{4!} f''''(x)(\xi_-) \right] + A_2 f(x) + A_3 \left[ f(x) + h f'(x) + \frac{h^2}{2!} f''(x) + \frac{h^3}{3!} f'''(x) + \frac{h^4}{4!} f''''(x)(\xi_+) \right], \]  

(18)

factoring, the above expression reduces to

\[ (A_1 + A_2 + A_3) f(x) + (-A_1 + A_3) hf'(x) + (A_1 + A - 3) \frac{h^2}{2!} f''(x) + \]  

\[ (-A_1 + A_3) \frac{h^3}{3!} f'''(x) + [A_1 f'''(x)(\xi_+) + A_3 f'''(x)(\xi_-)] \frac{h^4}{4!} \]  

(19)

the terms of equation (19) is equal to the equation (14) to obtain the following system of equations

\[ A_1 + A_2 + A_3 = 0 \]  

\[ A_1 - A_3 = 0 \]  

\[ A_1 + A_3 = \frac{2}{h^2} \]  

(20)

These expressions can be expressed in matrix

\[
\begin{bmatrix}
1 & 1 & 1 & 0 \\
-1 & 0 & 1 & 0 \\
1 & 0 & 1 & \frac{2}{h^2}
\end{bmatrix}
\]  

(21)
Developing the matrix in (21) we find the coefficients (weights) \( A_1 = \frac{1}{h^2}, A_2 = -\frac{2}{h^2} \) and \( A_3 = \frac{1}{h^2} \). This way you can find the coefficients (weights) for each operator of \( p \)-order. The operator (7) can now be written using the calculated coefficients.

\[
\frac{\partial^2 P}{\partial x^2} = \frac{1}{\Delta h^2} \left( P_{i-1,j}^{n} - 2P_{i,j}^{n} + P_{i+1,j}^{n} \right). \tag{22}
\]

For a scheme of order \( q \),

\[
D_n^{(2q)} P = \left[ \sum_{s=-q}^{q} \frac{A_s}{(\Delta x)^2} P_{i-(s+1),j}^{(n)} \right], \tag{23}
\]

Similarly

\[
D_n^{(2q)} P = \left[ \sum_{s=-q}^{q} \frac{A_s}{(\Delta x)^2} P_{i,j-(s+1)}^{(n)} \right], \tag{24}
\]

so that the Laplacian is written as shown in equation (4). To calculate the \( A_s \), through the expansion by Taylor series is:

\[
f(x - qh) = f(x) - \frac{qh}{1!} f'(x) + \frac{(qh)^2}{2!} f''(x) - \frac{(qh)^3}{3!} f'''(x) + \frac{(qh)^4}{4!} f''''(x)(\xi_-) + \ldots (-1)^q (qh)^{2q} (2q)! \ldots \tag{25}
\]

\[
f(x + qh) = f(x) + \frac{qh}{1!} f'(x) + \frac{(qh)^2}{2!} f''(x) + \frac{(qh)^3}{3!} f'''(x) + \frac{(qh)^4}{4!} f''''(x)(\xi_+) + \ldots (-1)^q (qh)^{2q} (2q)! \ldots \tag{26}
\]

replacing (24) and (25) in (23)

\[
D_n^{(2q)} f(x) = A_1 \left[ f(x) - qhf'(x) + \frac{(qh)^2}{2!} f''(x) - \frac{(qh)^3}{3!} f'''(x) + \frac{(qh)^4}{4!} f''''(x)(\xi_-) \right] + A_2 f(x) + \\
A_3 \left[ f(x) + qhf'(x) + \frac{(qh)^2}{2!} f''(x) + \frac{(qh)^3}{3!} f'''(x) + \frac{(qh)^4}{4!} f''''(x)(\xi_+) \right]. \tag{27}
\]

Sorting the equation (26) in powers of \( qh \), ie, \( qh^0, qh^1, qh^2, \ldots, qh^{2q} \), a system of linear equations where the unknowns are the coefficients obtained \( A_s \). Once these coefficients are calculated numerically calculated the Laplacian given by equation (4).
4 NUMERICAL SCHEMES IN FINITE DIFFERENCES

The numerical schemes implemented in this paper are based on the expression (2) using the expression Laplacian shown in equation (4) for different orders of $p$.

For $q = 1$ we obtain the numerical finite difference scheme using a second-order approximation in time and space $O(\Delta t^2, \Delta h^2)$.

$$P_{i,j}^{n+1} = 2P_{i,j}^{(n)} - P_{i,j}^{n-1} + G^2\sum_{s=1}^{3} A_s \left\{ \frac{1}{\Delta x^2} \left( P_{i-(p-s+1),j}^{n} \right) + \frac{1}{\Delta z^2} \left( P_{i,j-(p-s+1)}^{n} \right) \right\},$$  \hspace{1cm} (28)

with $A_1 = 1$, $A_2 = -2$ and $A_3 = 1$

where $G = \frac{c\Delta t}{\Delta h}$, and $\Delta h$, and is the size of the mesh, it is clear that a square mesh is taken that is: $\Delta h = \Delta x = \Delta z$, which is considered stable for the following condition deduced by considering the stability condition shown above in (5)

$$G \leq \frac{1}{\sqrt{2}}.$$  \hspace{1cm} (29)

For $q = 2$, we obtain the numerical finite difference scheme using a fourth-order approximation in space and second-order in time $O(\Delta t^2, \Delta h^4)$.

$$P_{i,j}^{n+1} = 2P_{i,j}^{(n)} - P_{i,j}^{n-1} + G^2\sum_{s=1}^{5} A_s \left\{ \frac{1}{\Delta x^2} \left( P_{i-(p-s+1),j}^{n} \right) + \frac{1}{\Delta z^2} \left( P_{i,j-(p-s+1)}^{n} \right) \right\},$$  \hspace{1cm} (30)

with $A_1 = -\frac{1}{12}$, $A_2 = \frac{4}{3}$, $A_3 = -\frac{5}{2}$, $A_4 = \frac{4}{3}$ and $A_5 = \frac{1}{12}$

where the stability condition by symmetry shown above in (5) is reduced to

$$G \leq \sqrt{3}.$$  \hspace{1cm} (31)

For $q = 3$ we obtain the numerical finite difference scheme using a sixth-order approximation in space and second-order in time $O(\Delta t^2, \Delta h^6)$. Laplacian whose coefficients are calculated and obtaining the following numerical scheme

$$P_{i,j}^{n+1} = 2P_{i,j}^{(n)} - P_{i,j}^{n-1} + G^2\sum_{s=1}^{7} A_s \left\{ \frac{1}{\Delta x^2} \left( P_{i-(p-s+1),j}^{n} \right) + \frac{1}{\Delta z^2} \left( P_{i,j-(p-s+1)}^{n} \right) \right\},$$  \hspace{1cm} (32)

with $A_1 = -\frac{1}{90}$, $A_2 = -\frac{3}{20}$, $A_3 = \frac{3}{2}$, $A_4 = -\frac{49}{18}$ and $A_5 = \frac{3}{2}$, $A_6 = \frac{3}{20}$ and $A_7 = \frac{1}{90}$ here,
to preserve the stability of the numerical scheme (see eq 5).

For \( q = 4 \) an Finite-difference scheme of eighth order \( O(\Delta t^2, \Delta h^8) \) To consider an eighth order Finite-difference operators, is here necessary to take 9 grid points.

\[
P_{i,j}^{n+1} = 2P_{i,j}^{(n)} - P_{i,j}^{n-1} + G^2 \sum_{s=1}^{9} A_s \left\{ \frac{1}{\Delta x^2} (P_{i-(p-s+1),j}^n) + \frac{1}{\Delta z^2} (P_{i,j-(p-s+1)}^n) \right\},
\]

(34)

with \( A_1 = -\frac{1}{315}, A_2 = \frac{8}{315}, A_3 = -\frac{1}{5}, A_4 = \frac{8}{5}, A_5 = -\frac{205}{72}, A_6 = \frac{8}{5} \) and \( A_7 = -\frac{1}{5}, A_8 = \frac{8}{315} \) and \( A_9 = -\frac{1}{560} \) here,

\[
G \leq \sqrt{\frac{315}{1024}},
\]

(35)

to preserve the stability of the numerical. Finally for \( q = 5 \) the Finite-difference scheme of tenth order \( O(\Delta t^2, \Delta h^{10}) \), the scheme is

\[
P_{i,j}^{n+1} = 2P_{i,j}^{(n)} - P_{i,j}^{n-1} + G^2 \sum_{s=1}^{11} A_s \left\{ \frac{1}{\Delta x^2} (P_{i-(p-s+1),j}^n) + \frac{1}{\Delta z^2} (P_{i,j-(p-s+1)}^n) \right\},
\]

(36)

with \( A_1 = -\frac{1}{3150}, A_2 = -\frac{5}{1008}, A_3 = -\frac{5}{126}, A_4 = \frac{5}{21}, A_5 = -\frac{5}{3}, A_6 = \frac{1086}{371} \), by symmetry \( A_7 = A_5, A_8 = A_4, A_9 = A_3, A_{10} = A_2 \) and \( A_{11} = A_1 \) where

\[
G \leq \sqrt{\frac{4}{9.75}},
\]

(37)

to preserve the stability of the numerical. Numerical modeling using the scheme described in (23), (25), (27), (29) y (31) is used later to observe the compressional wave field over time, and this scheme is compared.
5 NUMERICAL MODELING AND RESULTS

For the first numerical example, the velocity model is 1500 m/s, the number of samples in the direction of depth in both the horizontal are 512 samples, the spacing is 10 meters, that gives us a length of 5120 meters in both directions also the seismic source is located in the middle of the model \((x_s, z_s) = (2560m, 2560m)\) The wavefield in 1000 time steps is observed, considering a time step according to (5), for a scheme shown in (38) see Figure 1a , for a scheme shown in (30) see Figure 1b, for a scheme shown in (32) see Figure 1c, for a scheme shown in (34) see Figure 1d, and for a scheme shown in (36) see Figure 1e

![Wavefronts using different schemes](image.png)

**Figure 1**: a) Wavefront using a scheme \(O(\Delta t^2, \Delta h^2)\), b) using a scheme \(O(\Delta t^2, \Delta h^4)\), c) using a scheme \(O(\Delta t^2, \Delta h^6)\), d) using a scheme \(O(\Delta t^2, \Delta h^8)\) and e) for Scheme \(O(\Delta t^2, \Delta h^{10})\)
The second numerical experiment is based on a model of 3 horizontal layers, the seismic velocities in the model are respectively 3000, 3500, 4500 and 5000 m/s was obtained the graph of $x$ vs. $t$ for the implemented numerical schemes where the source and receivers were located near surface. In Figure 2 the seismograms observed numerical schemes for different model in horizontal layers, for a scheme shown in (38) see Figure 2a, for a scheme shown in (30) see Figure 2b, for a scheme shown in (32) see Figure 2c, for a scheme shown in (34) see Figure 2d, and for a scheme shown in (36) see Figure 2e. The numerical modeling parameters are the same as the homogeneous model but the mesh size used in all schemes is 20 meters, all seismograms border problems are observed in the numerical modeling.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{seismograms}
\caption{a) Seismogram using a scheme $O(\Delta t^2, \Delta h^2)$, b) using a scheme $O(\Delta t^2, \Delta h^4)$, c) using a scheme $O(\Delta t^2, \Delta h^6)$, d) using a scheme $O(\Delta t^2, \Delta h^8)$, and e) for Scheme $O(\Delta t^2, \Delta h^{10})$}
\end{figure}
Finally the numerical dispersion graphs are shown for the phase velocity and the group velocity using the scheme of second order in time and space (see Figure 3a and 3b respectively). curves in these figures are obtained for different values of $p$. Here $p = \frac{c_0 \Delta t}{\Delta x}$ is a function of the speed and spacing of the computational mesh., the expressions for these curves are:

$$v_g \frac{c_0}{c_0} = \frac{\cos(\pi \Delta x/\lambda)}{\sqrt{1 - p^2 \sin^2(\pi \Delta x/\lambda)}},$$  \hspace{1cm} (38)

$$v_f \frac{c_0}{c_0} = \frac{4\lambda}{2p\pi \Delta x} \arcsin\left(p \sin\left(2\pi \Delta x/4\lambda\right)\right)$$ \hspace{1cm} (39)

![Figure 3: a) Phase speeds and b) group velocity normalized using second order scheme](image)

6 CONCLUSIONS

It showed that the method of undetermined coefficients based on Taylor Expansion helps operators to calculate the second derivative for any order. He showed in this paper that the numerical dispersion error decreases when the order finite difference scheme increases. An order of six or eight parallel order scheme is sufficient to give good results. Numerical modeling algorithms were developed using stable finite difference schemes for different orders of approximation These numerical algorithms can be used for RTM migration in the time domain to calculate the wave field propagated and retro propagated

7 REFERENCES

finite-difference approach, Geophysics 41: 2-27.

from bodies of arbitrary shape, Geophysics 41: 145-150.

equation, Geophysics 51: 5466.


the scalar and elastic wave equation, Geophysics 49: 533-549.

finite difference method, Geophysics 51: 889-901.

fracture-induced velocity and attenuation anisotropy, Geophysical Journal International
191: 1179-1191.

in a linear viscoacoustic medium, Geophysical Journal International 93: 393–401.


Planewave theory and numerical simulation The Journal of the Acoustical Society of
America 99: 2655–2666.

Homogeneous Linear DEs with Constant Coefficients. Alabama Journal of Mathematics,

Geophysics, 48: 1514–1524.

(SBGF) Pag 141.

of finite-difference waveequation computations, 64(3), 967-969.

[16] Liu, Y., and Sen, M.K (2009) Advanced finite-difference methods for seismic mod-
eling, Geohorizons 14(2) :5–16
COMPUTATIONAL METHOD FOR INTERACTIONS BETWEEN COMPRESSIBLE FLUIDS AND SOLIDS WITH THERMAL CONDUCTIVITY

D. Toriu\textsuperscript{1} and S. Ushijima\textsuperscript{2}

\textsuperscript{1}Department of Civil and Earth Resources Engineering, Graduate School of Engineering, Kyoto University, Kyoto, Japan, 615-8540
toriu.daisuke.57x@st.kyoto-u.ac.jp

\textsuperscript{2}Academic Center for Computing and Media Studies (ACCMS), Kyoto University, Kyoto, Japan, 606-8501
ushijima@media.kyoto-u.ac.jp

Key words: Fluid-Solid Interaction, Heat Transfer, Compressible Fluid

Abstract. In this paper, we propose the computational method for mechanical and thermal interactions between compressible fluids and solids. The multi-phase field consisting of compressible fluids and solids is assumed to be the one fluid, and averaged governing equations are solved based on the fractional time-step method. The present method enables us to calculate the compressible flow around solids and the heat transfer between fluids and solids using simple orthogonal structured grids.

In order to validate the accuracy of the present method, the compressible flow around a wedge was calculated. The calculated oblique shock angle was compared with the theoretical result. In addition, the numerical experiments for the natural convection and the heat conduction around and in a horizontal circular pipe were conducted. Through these calculations, the validity of the calculated temperature distributions and the effect of the thermal conductivity were discussed.

1 INTRODUCTION

There are many engineering subjects that involve mechanical and thermal interactions between fluids and solids. In some cases, the compressibility of fluids caused by temperature and pressure changes should be estimated accurately, e.g. the helium leakage from the spent fuel storage container [1].

Yamamoto et al.\textsuperscript{[2]} proposed the preconditioning method for the problem of the mechanical and thermal coupling between fluids and solids using compressible Navier-Stokes...
equations on the boundary-fitted coordinate system. This method can calculate not only the compressible flow around solids but also the natural convection in fluids and the heat conduction in solids simultaneously taking into account the compressivity of fluids. On the other hand, the method using simple orthogonal grids is also useful in applications for complex geometries, because the mesh generation and the numerical treatment are simple in this method.

In this paper, we investigate the numerical method for the mechanical and thermal interactions between compressible fluids and solids based on the multi-phase model [3]. The multi-phase field is treated as the one fluid which has spatially different properties, and the averaged governing equations are solved. Therefore, the present method enables us to calculate flows around solids and heat transfer between fluids and solids using simple orthogonal structured grids. In order to confirm the applicability of the present method, we calculated the compressible flow around a wedge and the natural convection and the heat conduction around and in a horizontal circular pipe.

2 COMPUTATIONAL METHOD

2.1 Governing Equations

In this study, the multi-phase field consisting of compressible fluids and solids is treated as the one fluid which has spatially different properties. Governing equations for the multi-phase field are given by following averaged mass conservation equation in Eulerian description, momentum equations and energy equation.

\[ \frac{\partial \rho}{\partial t} + \frac{\partial (\rho u_i)}{\partial x_i} = 0 \]  
\[ \frac{\partial (\rho u_i)}{\partial t} + \frac{\partial (\rho u_i u_j)}{\partial x_j} = \frac{\partial \sigma_{ij}}{\partial x_j} + \rho f_i \]  
\[ \frac{\partial (\rho e)}{\partial t} + \frac{\partial (\rho e u_j)}{\partial x_j} = \sigma_{ij} \frac{\partial u_i}{\partial x_j} - \frac{\partial q_j}{\partial x_j} \]

Here, \( t, x_i \) and \( f_i \) are time, \( i \)-th component of two-dimensional orthogonal coordinates and acceleration of external force in \( x_i \) direction respectively. In Eqs.(1), (2) and (3), averaged density \( \rho \), velocity components \( u_i \), stress \( \sigma_{ij} \), internal energy \( e \) and heat flux components \( q_j \) are defined as

\[ \rho = \sum_k \phi_k \rho_k \]  
\[ u_i = \frac{1}{\rho} \sum_k \rho_k u_{k,i} \]
\[ \sigma_{ij} = \sum_k \phi_k \sigma_{k,ij} \quad (6) \]

\[ e = \frac{1}{\rho} \sum_k \rho_k e_k \quad (7) \]

\[ q_j = \sum_k \phi_k q_{k,j} \quad (8) \]

where subscript \( k \) represents phase \( k \) and \( \phi_k \) is volume fraction of phase \( k \) to the control volume. Averaged stress \( \sigma_{ij} \) and heat flux components \( q_j \) are approximately given as follows:

\[ \sigma_{ij} = \sum_k \phi_k \sigma_{k,ij} \approx -p \delta_{ij} + \tau_{ij} \quad (9) \]

\[ q_j = \sum_k \phi_k q_{k,j} \approx -\lambda \frac{\partial T}{\partial x_j} \quad (10) \]

Here, Eq.(9) is derived from the assumption that the solid is stationary and undeformable. Averaged viscous stress \( \tau_{ij} \), thermal conductivity \( \lambda \) and temperature \( T \) are given by

\[ \tau_{ij} = \mu \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) - \frac{2}{3} \mu \frac{\partial u_m}{\partial x_m} \delta_{ij} \quad (11) \]

\[ \lambda = \sum_k \phi_k \lambda_k \quad (12) \]

\[ T = \sum_k \phi_k T_k \quad (13) \]

where \( \mu \) is the viscosity coefficient of fluid and \( \delta_{ij} \) is Kronecker delta respectively. In addition, pressure \( p \) used in Eq.(9) is directly obtained from the pressure calculation.

The fluid is assumed to be the ideal gas. The equation of state is given by

\[ p_g = (\gamma_g - 1) \rho_g e_g \quad (14) \]

where subscript \( g \) represents the gas and \( \gamma \) is the specific heat ratio. The relationship between internal energy \( e \) and temperature \( T \) is given as follows:

\[ e = C_V T \quad (15) \]

Here, averaged specific heat at constant volume and pressure, \( C_V \) and \( C_P \), are defined as

\[ C_V = \phi_g C_{Vf} + \phi_s C_{Ps} \quad (16) \]

\[ C_P = \phi_g C_{Pf} + \phi_s C_{Ps} \quad (17) \]
2.2 Numerical procedures

We adopted the fractional time-step method for the time integration of the equations. In the present method, numerical procedures are divided into three stages, convection, diffusion and pressure calculation stage, as shown in Fig.1. Here, the variables $Q$ after each stage are represented as $Q^*$, $Q^{**}$ and $Q^{n+1}$ respectively. The convection, diffusion and pressure terms are calculated considering interactions between fluids and solids in each stage. For the spatial discretization of the governing equations in each stage, we use the finite volume method (FVM) on the orthogonal collocated grid system.

In the convection calculation stage, variables $\rho^*$, $(\rho u_i)^*$ and $(\rho e)^*$ are calculated by following convection equations written in the conservative form.

$$\frac{\rho^* - \rho^n}{\Delta t} + \frac{\partial (\rho^n u^n_j)}{\partial x_j} = 0 \quad (18)$$

$$\frac{(\rho u_i)^* - (\rho u_i)^n}{\Delta t} + \frac{\partial \{(\rho u^n)_i u^n_j\}}{\partial x_j} = 0 \quad (19)$$

$$\frac{(\rho e)^* - (\rho e)^n}{\Delta t} + \frac{\partial \{(\rho e)_i u^n_j\}}{\partial x_j} = 0 \quad (20)$$

For the convection terms, third-order MUSCL TVD scheme [4] is used. Since $\rho^*$ is estimated by the equation written in conservative form as shown in Eq.(18), the mass conservative law is sufficiently satisfied in this study.

The velocity components $u_i^*$ are given by the following equations.

$$u_i^* = \frac{1}{\rho^*} [\phi_g (\rho u_i)^* + \phi_s \rho_s u_{s,i}] \quad (21)$$

Here, $\rho^*$ is constant and $u_{s,i}$ are 0 in this study. The temperature $T^*$ and the pressure $p^*$ are calculated by

$$T^* = \frac{1}{\rho^* C_V} (\rho e)^* \quad (22)$$

$$p^* = (\gamma - 1) \rho^* e^* \quad (23)$$

![Figure 1: Flow chart of numerical procedures](image-url)
The variable changes caused by the diffusion terms are estimated in the diffusion calculation stage. These changes are written as

\[
\frac{\rho^{**} - \rho^*}{\Delta t} = 0 \tag{24}
\]

\[
\frac{(\rho u_i)^{**} - (\rho u_i)^*}{\Delta t} = \frac{\partial \tau_{ij}^*}{\partial x_j} \tag{25}
\]

\[
\frac{(\rho e)^{**} - (\rho e)^*}{\Delta t} = \tau_{ij}^* \frac{\partial u_i^*}{\partial x_j} - \frac{\partial q_j^*}{\partial x_j} \tag{26}
\]

As shown in Eq.(24), the density is constant in this stage (i.e. \(\rho^{**} = \rho^*\)). The velocity components \(u_i^{**}\) are given by

\[
u_i^{**} = \frac{1}{\rho^{**}} [\phi_g (\rho u_i)^{**} + \phi_s \rho_s u_{s,i}] \tag{27}
\]

The temperature \(T^{**}\) and the pressure \(p^{**}\) are calculated by following equations based on TCUP method [5].

\[
\frac{\rho^* C_V (T^{**} - T^*)}{\Delta t} = \frac{\partial (\tau_{ij}^* u_i^*)}{\partial x_j} - \frac{\partial q_j^*}{\partial x_j} - \frac{\rho^* (u_i^{**} - u_i^*)}{2\Delta t} \tag{28}
\]

\[
p^{**} - p^* = \phi_g \frac{\gamma - 1}{\gamma} \frac{\rho^{**} C_p}{\rho^* C_p \mu_J + 1} (T^{**} - T^*) \tag{29}
\]

where \(\mu_J\) is Joule-Thomson coefficient. For the constant temperature in the solids, \(T^{**}\) is calculated by the following equation instead of Eq.(28).

\[
T^{**} = \phi_g \left( T^* + \frac{\Delta t}{\rho^* C_V} \Theta^{**} \right) + \phi_s T_s \tag{30}
\]

Here, \(\Theta^{**}\) represents the right hand side of Eq.(28).

In the pressure calculation stage, we solve the pressure equation used in CCUP method [6] and TCUP method [5] as

\[
\frac{1}{\rho^{**} (C_s^{**})^2} \frac{p^{n+1} - p^{**}}{\Delta t} = -\frac{\partial}{\partial x_i} \left( -\frac{1}{\rho^{**}} \frac{\partial p^{n+1}}{\partial x_i} \Delta t + u_i^{**} \right) \tag{31}
\]

where \(C_s\) is the sound speed given by \(C_s = \sqrt{\gamma g \rho / \rho}\). As shown in Eq.(31), the pressure is calculated implicitly in this study. This enables us to adopt the large time increment \(\Delta t\).

The numerical stability in the calculation method, which exactly estimates the density change caused by the pressure, is limited by the Courant number for the sound speed as follows:
\[ C = \max_i \left( \frac{|u_i| + C_s \Delta t}{\Delta x_i} \right) \]  

(32)

where \( C \) is the Courant number and \( \Delta x_i \) is the grid size in \( i \)-th direction. The Courant number should be less than 1 (\( C < 1 \)) for the numerical stability. In the calculation of low speed flows, e.g. natural convection, this limitation leads to the increase in the number of iteration steps because the sound speed is much larger than the flow velocity. In the present method, we adopted the implicit algorithm for pressure calculation in order to use large \( \Delta t \) in the calculation of the natural convection.

The next time step velocity components \( u_i^{n+1} \) and the internal energy \( e^{n+1} \) are given by

\[ u_i^{n+1} = \frac{1}{\rho^{**}} \left[ \phi_g (\rho u_i)^{n+1} + \phi_s \rho_s u_{s,i} \right] \]  

(33)

\[ \frac{(\rho e)^{n+1} - (\rho e)^{**}}{\Delta t} = -p^{n+1} \frac{\partial u_i^{n+1}}{\partial x_i} \]  

(34)

In addition, we use directly \( \rho^{**} \) as \( \rho^{n+1} \) (i.e. \( \rho^{n+1} = \rho^{**} = \rho^* \)).

3 APPLICATIONS

3.1 Compressible Flow around a Wedge

The present method was applied to the two-dimensional compressible flow around a wedge. The calculation area and the wedge are shown in Fig.2. In Fig.2, \( L_1 = 0.2 \) [m], \( L_2 = 0.3 \) [m], and the half angle of the wedge \( \theta_c \) is 20 [deg]. The fluid is inviscid and nonconductive air (\( \gamma_g = 1.4 \)). On left and right hand side boundaries, inlet and free outlet conditions are imposed. The inlet velocities \( u_{i,in} \) are constant, \( u_{1,in} / C_s = 3 \) (Mach number is 3) and \( u_{2,in} = 0 \) [m/s]. In addition, the velocity components in the initial condition \( u_{i,0} \) were \( u_{i,0} = u_{i,in} \). The time increment and the computational mesh size are \( \Delta t = 5.0 \times 10^{-8} \) [s] and 600 × 900 respectively. The computation was parallelized by OpenMP [7] in order to reduce the computational time and conducted on the supercomputer Cray-XE6 of ACCMS, Kyoto University, with 8 threads.

The calculated density isolines in the steady state are shown in Fig.3. The oblique shock attached to the tip of the wedge was calculated stably by the present method. Through this calculation, the Courant number \( C \) was about 0.21. In Fig.3, \( \theta_s \) is the oblique shock angle obtained theoretically. Between \( \theta_s \), \( \theta_c \), and Mach number \( M \), following relational expression is established [8].

\[ \tan \theta_c = 2 \cot \theta_s \left[ \frac{M^2 \sin^2 \theta_s - 1}{2 + M^2 (\gamma_g + \cos 2\theta_s)} \right] \]  

(35)
where $\theta_s = 37.76 \text{ [deg]}$ in this calculation condition. As shown in Fig.3, the calculated oblique shock angle was good agreement with the theoretical result. This shows that the present method enables us to predict reasonably the compressible flow around the solid.

4 Natural Convection and Heat Conduction around and in a Horizontal Circular Pipe

The present method was applied to the two-dimensional natural convection and heat conduction around and in a horizontal circular pipe reported by Yamamoto et al.[2]. Through this numerical experiment, the applicability of the present method to the heat transfer problem between fluids and solids is discussed.

The horizontal circular pipe is located in the calculation area as shown in Fig.4. The inner and the outer radius of the pipe are $R_1 = 8.9 \times 10^{-3} \text{ [m]}$ and $R_2 = 1.78 \times 10^{-2} \text{ [m]}$ respectively. The lengths of the calculation area are sufficiently large compared with $R_2$, $L_1 = 54R_2$ and $L_2 = 72R_2$. The inner wall of the pipe and the boundaries of the calculation area (non-slip wall) are fixed at $T_h = 325.5 \text{ [K]}$ and $T_c = 293 \text{ [K]}$ respectively.

The working gas is air, Prandtl number $Pr = 0.705$ and $\gamma_g = 1.4$. In the initial state, Rayleigh number is $Ra = 1.02 \times 10^5$. Here, $Ra$ is given by

$$Ra = \frac{g \beta_0 (2R_2)^3 (T_h - T_c)}{\nu_0 \alpha_0} \quad (36)$$
where \( g \), \( \beta \), \( \nu \), and \( \alpha \) are the gravitational acceleration, the coefficient of the volumetric thermal expansion, the coefficient of the kinematic viscosity, and the thermal diffusivity respectively. The subscript 0 represents the value in the initial state.

The body of the pipe has the thermal conductivity \( \lambda_s \). In this application, three different ratio of the thermal conductivities \( K = 5, 10, \) and \( 20 \), where \( K = \lambda_s/\lambda_g \), are taken into account. The density and the specific heat of the pipe, \( \rho_s \) and \( C_{Ps} \), are equal to those of the air. The computational mesh size and the time increment are \( 2160 \times 2880 \) and \( \Delta t = 5.0 \times 10^{-5} \) [s] in all cases. We employed the parallel computational method based on the 2D domain decomposition with flat MPI [9]. The computation was executed on the supercomputer Cray-XC30 of ACCMS, Kyoto University with \( 9 \times 12 \) processes. Through calculations for \( K = 5, 10, \) and \( 20 \), the Courant numbers for the sound speed defined as Eq.(32) were about \( C = 40 \). In the present method, since the pressure is calculated implicitly as shown in Eq.(31), we can adopt large \( \Delta t \) over \( C = 1 \) in the calculation of the low speed flow.

Figure 5 and 6 show calculated temperature contours and distributions around the outer wall of the pipe and in the body of the pipe. Here, distance from the outer wall of the pipe \( Y \) and angle \( \theta \) are defined as Fig.7. Dimensionless numbers \( \phi \) and \( Y^* \) used in Fig.6 are given by \( \phi = (T - T_c)/(T_h - T_c) \) and \( Y^* = Ra^{1/4}Y/(2R_2) \). As shown in Fig.5 and 6, the calculated temperature contours and distributions in the body of the pipe and the air are smoothly connected on the outer wall of the pipe. In the body of the pipe, the temperature of the upper part is higher than others in all cases. This nonuniformity of the temperature distribution in the body of the pipe is caused by the heat release on the outer wall of the pipe due to the natural convection. In addition, the increase of \( K \) caused the increase of the temperature in the body of the pipe as shown in Fig.5 and 6. The temperature of the air around the pipe was influenced by the differences of the
temperature on the outer wall of the pipe.

The same tendencies, the nonuniformity of the temperature and the influence of thermal conductivity of the pipe, have been reported by Yamamoto et al. [2]. In their study, the O-type grids fitting to the inner and the outer wall of the pipe were used. From these results, it was shown that the present method enables us to calculate reasonably and stably the thermal interactions between fluids and solids using the averaged governing equations for the multi-phase field.

![Temperature contours around and in the pipe](image1)

(a) $K = 5$
(b) $K = 10$
(c) $K = 20$

**Figure 5**: Temperature contours around and in the pipe

![Temperature distributions at θ = 90 and 180 deg](image2)

(a) $\theta = 90$ [deg]
(b) $\theta = 180$ [deg]

**Figure 6**: Temperature distributions at $\theta = 90$ and 180 [deg]
5 CONCLUSIONS

To calculate mechanical and thermal interactions between compressible fluids and solids, the multiphase computational method was investigated. In the present method, the averaged governing equations for the multi-phase field are solved based on the fractional time-step method. The present method enables us to reasonably and stably calculate the compressible flow around solids and the heat transfer between fluids and solids on simple orthogonal grids.

As a result of the computations for the compressible flow around the wedge, the calculated oblique shock angle was in good agreement with the theoretical result. It was also confirmed that the heat transfer between the fluid and solid are reasonably predicted taking into account the effect of the thermal conductivity of the solid.

REFERENCES


NUMERICAL SIMULATION OF THE WEISSENBERG EFFECT

RAFAEL A. FIGUEIREDO¹, CASSIO M. OISHI² AND JOSÉ A. CUMINATO³

¹Instituto de Ciências Matemáticas e Computação (ICMC), Universidade de São Paulo
Av. do Trabalhador São-Carlense, 400, São Carlos - SP, Brazil
E-mail: rafigueiredo22@gmail.com

²Faculdade de Ciência e Tecnologia (FCT), Universidade Estadual Paulista Julio de Mesquita Filho
Roberto Simonsen, 305, Presidente Prudente - SP, Brazil
E-mail: cassiooishi@gmail.com

³Instituto de Ciências Matemáticas e Computação (ICMC), Universidade de São Paulo
Av. do Trabalhador São-Carlense, 400, São Carlos - SP, Brazil
E-mail: jacumina@icmc.usp.br

Key words: Weissenberg Effect, FDM, VOF, Viscoelastic Flow, Axisymmetric Flow

Abstract. Weissenberg effect, or Rob-Climbing experiment, is encountered in many important industrial applications involving mixing process. In this experiment we insert a rotating rod into a container with viscoelastic fluid. The behavior of the viscoelastic fluid is striking, the fluid moves in the opposite direction of the centrifugal force and climbs up the rod. This effect has been experimentally reproduced by many researchers. However, in the context of numerical results, there are few works dealing with a rigorous study about this effect. One of the main challenges imposed in this simulation is to provide numerical methods which achieve high elasticity rate and that represent the correct shape of the free surface. In this work is presented a finite difference scheme to simulate the Weissenberg effect. The numerical technique is based on the projection method to simulate viscoelastic two-phase flows. The conformation tensor is employed to solve the viscoelastic fluid model. The interface between the fluids is represented with the efficient least squares volume-of-fluid interface reconstruction algorithm (ELVIRA). A second-order operator-split method is used to solve the advection equation and obtain the volume fraction in the next step. In this work, we have performed simulations of the Weissenberg effect for the linear Phan-Thien-Tanner (LPTT) model.
1 INTRODUCTION

Viscoelastic fluid flows are present in many industrial applications such as in mixing processes. In this type of problem occurs an amazing effect called Rod-Climbing effect or Weissenberg effect [1]. This problem consists of a rotating rod into a container with viscoelastic fluid. Due to the normal stress difference, the fluid climbs at rod.

Many works performed experimental and theoretical predictions about the Weissenberg effect [2, 3, 4]. On the other hand, few works are found presenting numerical results about this effect. Debbaut and Hocq [5] presented numerical results with a finite element method (FEM) in axisymmetric coordinates. In their studies secondary flows were analyzed in Weissenberg effect problem and results were compared with experimental and analytical results. Also using FEM and axisymmetric coordinates, Luo [6] presented numerical results dealing with the UCM and PTT models. In addition, Luo [6] proposed an elaborate scheme of orthogonal trajectories to update structured and unstructured meshes showing good agreement between numerical and experimental results. Recently, Habla et al. [7] developed an algorithm for solving viscoelastic two phase flows in OpenFOAM code. In [7] the numerical results of the Weissenberg effect were obtained with the VOF method to represent the interface in the three dimensional case.

There are some numerical difficulties for performing this effect, for instance: the treatment of the interface between liquid and air involving large deformation, the calculation of the surface tension, the presence of high angular velocity, the stability of numerical methods to simulate highly elastic flows of viscoelastic fluids. Therefore, towards an accurate simulation of the Weissenberg effect can be considered a challenging in Computational Rheology.

In this work we present a technique to simulate the Weissenberg effect. The algorithm is based on a finite difference scheme for solving viscoelastic two phase flows. The axisymmetric coordinate system was adopted due to the geometry of the problem. The projection method was used to solve the governing equations and the VOF method was used to represent the interface. The methodology is assessed comparing our numerical results with those reported in the literature.

2 GOVERNING EQUATIONS

Assuming isothermal and incompressible fluids for two-phase flows the governing equations can be written as

\[ \nabla \cdot \mathbf{u} = 0, \]

\[ \rho \left( \frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} \right) = -\nabla p + \nabla \cdot \mathbf{\tau} + \nabla \cdot \left( \mu_s \left( \nabla \mathbf{u} + (\nabla \mathbf{u})^T \right) \right) + \mathbf{F}, \]

\[ \frac{\partial f}{\partial t} + \nabla \cdot (f \mathbf{u}) = 0. \]

In Eqs. (1) - (3), \( \mathbf{u} \) is the velocity field, \( t \) is time, \( \rho \) is the density, \( \mu_s \) is the solvent.
The viscoelastic model adopted to simulate the Weissenberg effect in this work is the linear Phan-Thien-Tanner (LPTT) [8]. This model satisfies the second normal stress differences, which is also responsible for the climbing effect of the fluid [1].

The constitutive equation of the LPTT model is given by

\[
\left(1 + \frac{\epsilon \lambda}{\mu_p} tr(T) \right) \tau + \lambda \Box \tau = 2 \mu_p D,
\]

where \(\Box\) represents the Gordon-Schowalter derivative

\[
\Box = \frac{\partial \tau}{\partial t} + \mathbf{u} \cdot \nabla \tau - (\nabla \mathbf{u})^T \cdot \tau - \tau \cdot \nabla \mathbf{u} + \xi (\tau \cdot D + D \cdot \tau),
\]

\(\mathbf{D} = \frac{1}{2} \left( \nabla \mathbf{u} + (\nabla \mathbf{u})^T \right)\) is the rate of deformation tensor, \(tr(\tau)\) is the trace of \(\tau\), \(\mu_p\) the polymeric viscosity, \(\lambda\) is the time relaxation time, and \(\epsilon\) and \(\xi\) are parameters related to the elongational viscosity and the ratio between second and first normal stress difference, respectively.

Due to the geometry of the Weissenberg effect problem, which is composed of a stick and a container cylinders, it is convenient to use cylindrical coordinate system for describing the mathematical model. For simplicity in this paper we assume the axisymmetric hypothesis.

Further details of operators and governing equations in cylindrical coordinates can be found in [1].

3 NUMERICAL METHOD

3.1 Temporal discretization and projection method for viscoelastic fluid flows

The Weissenberg effect problem and several other viscoelastic flows usually occur with low Reynolds numbers and high numbers of Weissenberg. This is a problem for explicit methods because the time step is defined as a function of the Reynolds number due to the parabolic stability restriction. To avoid this restriction in the temporal discretization we have applied an implicit discretization of the diffusive terms in Eq. (2) while the convective terms are explicitly computed. In particular, the numerical treatment used in the present paper for solving of the Navier-Stokes equations Eqs. (1) - (2) and the constitutive Eq. (4) is an adaptation of the projection method described in [9] in the solution of viscoelastic fluid flows. Summarily, we have:

\[
\nabla \cdot \mathbf{u}^{n+1} = 0,
\]

\[
\rho \left( \frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} + (\mathbf{u} \cdot \nabla \mathbf{u})^n \right) = -\nabla p^{n+1} + \nabla \cdot \left( \mu_s \left( \nabla \mathbf{u}^{n+1} + (\nabla \mathbf{u}^{n+1})^T \right) \right) + \nabla \cdot \mathbf{\tau}^{n} + \mathbf{F}^n.
\]
Due to the incompressibility constraint the numerical solution of the Eqs. (6) and (7) couples the velocity and pressure fields. To solve this problem the projection method was applied \cite{10}. The essence of the projection method is to calculate an intermediate velocity $\tilde{u}^n$, through Eq. (7) by replacing $p^{n+1}$ by $p^n$

$$
\rho \left( \frac{\tilde{u}^{n+1} - u^n}{\Delta t} + (u \cdot \nabla u)^n \right) = -\nabla p^n + \nabla \cdot \left( \mu_s \left( \nabla \tilde{u}^{n+1} + \left( \nabla \tilde{u}^{n+1} \right)^T \right) \right) + \nabla \cdot \tau^n + F^n. \quad (8)
$$

This strategy decouples the velocity and pressure fields, but it is necessary to correct the pressure value and calculate a velocity field that satisfies the Eq. (6). In this case the pressure correction is achieved by

$$
p^{n+1} = p^n + \frac{\psi^{n+1}}{\Delta t} - \mu_s \nabla^2 \psi^{n+1}, \quad (9)
$$

while the velocity field that satisfies the Eq. (6) is obtained by

$$
u^{n+1} = \tilde{u}^{n+1} - \frac{1}{\rho} \nabla \psi^{n+1}, \quad (10)
$$

where $\psi^{n+1}$ is calculated by the solution of the Poisson equation

$$
\nabla \cdot \left( \frac{1}{\rho} \nabla \psi^{n+1} \right) = \nabla \cdot \tilde{u}^{n+1}. \quad (11)
$$

After updating the velocity and pressure fields, remains update the polymeric tensor solving the constitutive equation and solve the convection equation of the interface to get the new position

$$
l \left[ \frac{\tau^{n+1} - \tau^n}{\Delta t} + u^{n+1} \cdot \nabla \tau^n - (\nabla u^{n+1})^T \cdot \tau^n - \tau^n \cdot \nabla u^{n+1} + \xi \left( \tau^n \cdot D^{n+1} + D^{n+1} \cdot \tau^n \right) \right] \\
+ \left( 1 + \frac{\epsilon \lambda}{\mu_p} \text{tr} (\tau^n) \right) \tau^n = 2 \mu_p D^{n+1}, \quad (12)
$$

$$
\frac{f^{n+1} - f^n}{\Delta t} = -\nabla \cdot \left( f^n u^{n+1} \right). \quad (13)
$$

### 3.2 Interface representation

The interface of the Weissenberg effect problem is represented by the volume of fluid method (VOF) \cite{11}. In the VOF method, a function $f \in [0, 1]$ is defined in each computational cell, known a volume fraction, and the fluid properties are interpolated at the interface cells as
\[ \chi = f \chi_1 + (1 - f) \chi_2, \quad (14) \]

where \( \chi \) can be any property of the fluid such as \( \rho, \mu_s, \mu_p, \lambda, \epsilon, \xi \) and the subscripts 1 and 2 represent the properties of each phase of the fluid.

For incompressible fluids the VOF interface is updated by solving the convection equation Eq. (13). In this work this evolution equation for the volume fraction is solved considering the geometrical procedure given by [12]. In summary, this advection scheme uses a second-order operator-split given by

\[
\tilde{f}_{i,j} = f^n_{i,j} + \frac{\Delta t}{r_i \Delta r} \left( \frac{F^n_{i,j} - F^n_{i+1,j}}{r_i} \right),
\]

\[
f^{n+1}_{i,j} = \tilde{f}_{i,j} + \frac{\Delta t}{\Delta z} \left( G^n_{i,j} - G^n_{i,j+1} \right) + \tilde{f}_{i,j} \left( u_{i,j+\frac{1}{2}} - u_{i,j-\frac{1}{2}} \right),
\]

where \( F^n_{i+\frac{1}{2},j} = (fu)_{i+\frac{1}{2},j} \) represents the flux of \( f \) through the right edge of the \( i, j \)th cell and \( G^n_{i,j+\frac{1}{2}} = (fv)_{i,j+\frac{1}{2}} \) represents the flux through the top edge of the \( i, j \)th cell. The operator-split performs the advection first in one direction and then in the other direction, preventing an amount of fluid to be advected twice causing problems in the numerical simulation.

To calculate the geometric advection is necessary to know the geometric of the interface. There are several techniques to build the interface in the VOF method, in which the most popular is the piecewise linear interface construction method (PLIC). In the PLIC method the interface is approximated by line segments in each cell of the computational domain, usually discontinuous at the borders of the cells.

One important feature in moving interface simulations is the numerical computation of the normal vector at the interface. Here, we have used the Least Squares VOF Interface Reconstruction Algorithm (ELVIRA) [12]. According to [12] the geometric advection algorithm combined with ELVIRA method to obtain interface information is very accurate and achieving good accuracy in simulations with smooth interface.

The idea of the ELVIRA is to calculate six candidates to the slope of the normal vector, \( \hat{m}_{i,j} \). In summary, by using backward, forward and central differences (denoted here as \( \hat{m}_b, \hat{m}_f \) and \( \hat{m}_c \), respectively) and considering the sum of rows and columns of a \( 3 \times 3 \) block centered in the \( i, j \)th cell, we have:
The slope that minimizes the error is chosen as the best approximation. The error is calculated by the $L^2$ norm on a $3 \times 3$ block centered in the $i, j$th cell as follows
\[
E(\tilde{m}_{i,j}) = \left[ \sum_{k,l=-1}^{1} \left( f_{i+k,j+l} - \tilde{f}_{i+k,j+l} \right)^2 \right]^{\frac{1}{2}},
\]
where $\tilde{f}$ is a linear approximation of $f$ in the $3 \times 3$ block with slope $\tilde{m}_{i,j}$, such that $\tilde{f}_{i,j} = f_{i,j}$.

### 3.3 Surface tension

In this paper the surface tension $F$ in the Eq. (2) is modeled by the continuum surface force (CSF) method [13]. This method considers the surface tension term as a surface force per unit area of the interface as
\[
F = \sigma \kappa n \delta_i,
\]
where $\sigma$ is the surface tension coefficient, $\kappa$ is the interface curvature, $n$ is the normal vector to the interface and $\delta_i$ is the $\delta$-function at the interface. In the Eq. (19) some terms are computed using the volume fraction, for instance $n = \nabla f / \| \nabla f \|$ and $\delta_i = \| \nabla f \|$, while the curvature is given by:
\[
\kappa = \kappa_1 + \kappa_2 = r_1^{-1} + r_2^{-1},
\]
where $r_1$ and $r_2$ are the radius of curvature. In axisymmetric coordinates $\kappa_1$ is calculated in the $r - z$ plane by the volume fraction according to Height Function (HF) method [14, 15], as described below.

In the HF method the curvature is computed by a finite difference scheme of a local height function. In two-dimensional axisymmetric case, three height functions are defined locally according to the largest component of normal vector. If $|n_z|_{i,j} > |n_r|_{i,j}$ the height functions are defined as the sum of the columns of a $7 \times 3$ block centered on the $i, j$th cell
\[ C_{i',j} = \Delta z \sum_{l=-3}^{3} f_{i',j+l}, \text{ for } i' = i - 1, i, i + 1. \]  \hspace{1cm} (21)

In an analogous manner, the height functions are calculated by the sum of the rows of a \(3 \times 7\) block when \(|n_r| > |n_z|\).

Once calculated the height functions \(C\), the curvature \(\kappa_1\) is defined in the HF method as

\[ \kappa_1 = \frac{C''}{(1 + (C')^2)^{3/2}}, \]  \hspace{1cm} (22)

where

\[ C' = \frac{C_{i+1,j} - C_{i-1,j}}{2\Delta r}, \]  \hspace{1cm} (23)

\[ C'' = \frac{C_{i+1,j} - 2C_{i,j} + C_{i-1,j}}{\Delta r^2}. \]  \hspace{1cm} (24)

The curvature \(\kappa_2\) is obtained through directly from trigonometry, i.e., \(\kappa_2 = n_r/r\), where \(n_r\) is the radial component of the unit normal vector.

4 NUMERICAL RESULTS OF THE WEISSENBERG EFFECT

As discussed earlier the Weissenberg effect or rod climbing occurs due to influence of normal stress differences in the variation of the pressure value in the radial direction [1]. The geometry of the problem consists of a rod of radius \(R\) with constant angular velocity \(\omega\), inserted in a cylindrical container of radius \(R_c\) and height \(H\) filled with viscoelastic fluid.

To verify the proposed methodology we have performed three simulations varying the angular velocity \(\omega = 1.7, 2.1\) and 2.7 rev/s, and fixing the geometric parameters, \(R = 0.635cm\), \(R_c = 10.8cm\) and \(H = 7.7cm\). Moreover, in the Weissenberg effect there are four dimensionless parameters that relate the properties of the fluid with the geometric parameters of the flow, the Weissenberg number \(Wi\), the Reynolds number \(Re\), the Stokes number \(St\) and the Capillary number \(Ca\), which are defined as

\[ Wi = \frac{\lambda \omega}{\mu}, \quad Re = \frac{\rho \omega R^2}{\mu}, \quad St = \frac{\mu \omega}{\rho g R}, \quad Ca = \frac{\mu \omega R}{\sigma}, \]

where \(\mu = \mu_s + \mu_p\) is the total viscosity and \(g\) is the gravity force.

In order to compare our results with those presented in the literature [2, 5, 6, 7], the following parameters were employed: \(\rho = 0.89g/cm^3\), \(\epsilon = 0\), \(\mu_s/\mu_p = 1/9\) and the other parameters were determined by the dimensionless values shown in Table 1. The parameter \(\xi\) is determined by a climbing constant \(\hat{\beta}\), where \(\hat{\beta}\) depends on the temperature and is obtained from experimental tests. According to [2, 6], at temperature 27\(^\circ\)C this constant
results in $\hat{\beta} = 0.86$ and at 26°C, we have $\hat{\beta} = 1.02$, resulting $\xi = 0.318$ and $\xi = 0.284$, respectively. The other phase was considered a newtonian fluid, with $\rho \to 0$ and $\mu_s \to 0$.

Table 1: Angular velocity and dimensionless parameter group used to verify the proposed methodology.

<table>
<thead>
<tr>
<th>$\omega$ (rev/s)</th>
<th>$Wi$</th>
<th>$Re$</th>
<th>$St$</th>
<th>$Ca$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.7</td>
<td>0.1944</td>
<td>0.0263</td>
<td>2.813</td>
<td>32.05</td>
</tr>
<tr>
<td>2.1</td>
<td>0.2401</td>
<td>0.0325</td>
<td>3.475</td>
<td>39.59</td>
</tr>
<tr>
<td>2.6</td>
<td>0.2973</td>
<td>0.0402</td>
<td>4.302</td>
<td>49.02</td>
</tr>
</tbody>
</table>

Figures 1-3 show the numerical results obtained with the proposed method varying the dimensionless parameter group of Table 1. In addition, in these figures, we have plotted the results presented by [2, 5, 6, 7]. These results were obtained with a uniform computational grid with spacing equal to 0.0395cm and $\Delta t = 5 \times 10^{-3}s$.

The climbing height of the fluid is measured by $h/r = (\tilde{h}(r) - H)/r$, where $\tilde{h}(r)$ is the height of the interface. In the numerical results, we can see that to $\xi = 0.284$ the fluid climbed about 3.5%, 4.8% and 6% more than $\xi = 0.318$, at $\omega = 1.7$, 2.1 and 2.6 rev/s, respectively. Similar results were obtained by [6, 7]. This difference in the climbing height in the results between the $\xi$ values, is consistent with the theory that the Weissenberg effect also depends on the second normal stress differences.

It can also be seen in Figs. 1-3 that as omega increases and consequently $Wi$ increases, the climbing height increases. Furthermore, the comparison between the numerical results obtained with the proposed methodology and the results reported in the literature, showed reasonable agreement. The difference between the results is due to the accuracy between the different numerical techniques.

5 CONCLUSIONS

In this paper we present a methodology to simulate the Weissenberg effect, also known as Rod-Climbing. The proposed method is based on a finite difference scheme, the projection method and the VOF technique to simulate axisymmetric viscoelastic two-phase flows.

The numerical results were obtained with the LPTT constitutive equation. The LPTT model was chosen in order to compare the results obtained with the proposed method, with the numerical results presented in the literature, which are few [5, 6, 7].

Three tests were performed from which we fixed the geometry parameters and varies the angular velocity. It was observed that when the angular velocity increases, the climbing height increases. The results showed reasonable agreement, indicating the good capability of the proposed method to simulate the Weissenberg effect.
Figure 1: Numerical results of the Weissenberg effect with $\omega = 1.7\, \text{rev/s}, \xi = 0.284, 0.318$, $Re = 0.0263$, $Wi = 0.1944$, $St = 2.813$ and $Ca = 32.05$.

Figure 2: Numerical results of the Weissenberg effect with $\omega = 2.1\, \text{rev/s}, \xi = 0.284, 0.318$, $Re = 0.0325$, $Wi = 0.2401$, $St = 3.475$ and $Ca = 39.59$.

6 ACKNOWLEDGMENTS

The authors acknowledge the financial support of the FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo, Grants N: Poti e Cassio, 2011/09194-7)
Figure 3: Numerical results of the Weissenberg effect with $\omega = 2.6\text{rev/s}$, $\xi = 0.284, 0.318$, $Re = 0.0402$, $Wi = 0.2973$, $St = 4.302$ and $Ca = 49.02$.

and CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico, Grant No. 309514/2013-4 and 473589/2013-3) which enabled this research to be performed through the projects linked to them.

REFERENCES


XFEM MODELLING OF HYDRAULIC FRACTURING

JAVIER B. LUCERO\textsuperscript{1}, MARIELA LUEGE\textsuperscript{2}, ANTONIO ORLANDO\textsuperscript{2} AND CELESTE TORRIJOS\textsuperscript{3}

\textsuperscript{1}CONICET. Inst. de Estructuras "Arturo M. Guzmán", FACET, UNT.
Av. Independencia 1800, S.M. de Tucumán, Argentina. 4000
jlucero@herrera.unt.edu.ar

\textsuperscript{2}CONICET. Inst. de Estructuras "Arturo M. Guzmán", FACET, UNT.
Av. Independencia 1800, S.M. de Tucumán, Argentina. 4000
mluege@herrera.unt.edu.ar, aorlando@herrera.unt.edu.ar

\textsuperscript{3}CONICET. LEMIT, UNLP.
52 Y 121, La Plata, Argentina. 1900.
celestetorrijos@hotmail.com

Key words: eXtended Finite Element Method, fractured porous media, fracking, hydraulic fracturing

Abstract. A computational model based on the extended finite element method (XFEM) is developed for hydraulic fracture analysis in porous media. The porous media equations are obtained by applying the balance equations of mass and momentum for both solid and fluid phase and assuming the Biot-Coussy poroelastic constitutive equations for modelling the behaviour of the solid phase, whereas the Darcy’s law is employed to describe the fluid transport within the intact part. As opposite, the flow through the crack is modelled by generalizing the Poiseuille’s law for viscous fluids to the flow between two parallel plates, and assuming the conductivity dependent on the dynamic viscosity and the crack opening displacement. The full coupled equations between the flow and the displacement field of the matrix are obtained by applying the mass balance equations within the discontinuity. The displacement field, pressure, and the discontinuities are approximated in the space using the XFEM, whereas the time discrete scheme is obtained by applying the backward Euler method. The resulting discrete nonlinear equations are solved by applying the Newton-Raphson’s method.

A model problem of hydraulic fracturing is analysed, by looking particularly at its behaviour near the crack-tip. Numerical experiments on the convergence with respect to mesh discretization and orientation are also presented.
1 INTRODUCTION

The modelling of fluids transport in fractured porous media is studied in many engineering applications such as hydraulic fracturing for petroleum engineering, permeability analysis for damage estimation and environmental engineering.

The hydraulic fracturing is a powerful technology that plays a vital role for the exploitation of non-conventional hydrocarbons sources. The numerical analysis is important in order to understand this very complex process which involves three different problems: the mechanical deformation caused by the fluid pressure that is exerted over the fracture faces, the fluid flux inside the fracture and the fracture propagation over a multiphase porous media where the phases interact among them. The full prediction of the behavior of the solid material deformation and of its interaction with a fluid flux is achieved by means of the full solution of the coupled problem.

The fluid transport in fractured porous media problems has been considered elsewhere: Boone and Ingraffea [1] presented a numerical procedure for the simulation of crack propagation caused by fluid injection in poroelastic materials, combining FEM with finite difference method; Schrefler [2] and Secchi et al. [3] modelled the cohesive fracture growth in full saturated porous media using FEM with adaptive meshes; In the last years, more appropriate methods for crack modelling in a continuum medium have emerged, such as the Embedded Finite Element Method (EFEM) Armero and Linder [4] and the eXtended Finite Element Method (XFEM) Moes et al. [5]. EFEM models the solid discontinuity using local dissipation through them. As long as, XFEM considers an enrichment kinematic, where the state variables are interpolated using regular functions and enrichment functions. These last are based on level set functions and they enable to describe a field using more general elements. XFEM has been used by de Borst et al.[6] and Mohammadnejad and Khoei [7] for porous media and by Lecampion [8] and Dahi-Teleghani et al. [9] for hydraulic fracture problems, and is the method used in this work.

In this paper a hydro-mechanical model for discontinuous saturated porous media is presented. The spatially discretization of the equations that governs the problem is obtained using XFEM by means of enrichment equations. The time integration is solved using the Euler’s implicit method. A Darcy’s flux is considered for the permeable matrix and a Poiselle’s one is used for the discontinuity. The propagation fracture condition is taken from Linear Elastic Fracture Mechanics (LEFM), using the Stress Intensity Factor concept. Finally, two numerical examples are presented to validate the procedure: The first are analyses the crack propagation caused by a mechanical action, whereas the second are considers the crack produced by fluid injection.

2 BALANCE EQUATIONS

2.1 Strong form of the problem

A saturated porous media $\Omega$ is constituted by a deformable solid phase (skeleton) and a porous space, which is completely occupied by the fluid phase. Assuming small defor-
mation, isothermal conditions, and neglecting gravitational terms, convective terms and accelerations, the momentum conservation equation can be written as

$$\nabla \cdot \sigma = 0 \quad \text{en } \Omega \quad (2.1)$$

where $\sigma$ is the Cauchy stress tensor. The mass balance equation of a porous media is

$$\frac{\partial \rho_\pi}{\partial t} + \rho_\pi \nabla \cdot \mathbf{v}_\pi = 0 \quad \text{en } \Omega \quad \text{con } \pi = s, f \quad (2.2)$$

where the subscripts $s$ and $f$ identifies the solid an fluid phases, respectively. By means of using the condition $n_s + n_f = 1$, where $n_s$ and $n_f$ represent the volume fractions of the solid and fluid phases. Introducing the expressions that relate the density changes of the solid mass with the volume change of the solid phase and the density change of the fluid mass with the fluid pressure (See Coussy [10] for more details) in (2.1), the following equation of mass conservation in $\Omega$ is obtained

$$\alpha \nabla \cdot \mathbf{v}_s + n_f \nabla \cdot (\mathbf{v}_f - \mathbf{v}_s) + \frac{1}{K_f} \frac{\partial p}{\partial t} = 0 \quad \text{en } \Omega \quad (2.3)$$

where $\mathbf{v}_s$ and $\mathbf{v}_f$ are the velocities of the solid an fluid phases respectively and $K_f$ is the compressibility modulus or Biot’s modulus. The problem description is completed by adding Dirichlet boundary conditions (for displacement and pressures) and Neumman conditions (for forces and fluid fluxes). The solid phase boundary conditions are written in terms of forces and displacements

$$\mathbf{n}_\Gamma \cdot \sigma = t_p \quad \text{en } \Gamma_t \quad \mathbf{u}_s = \mathbf{u}_p \quad \text{en } \Gamma_u \quad (2.4)$$

with $\Gamma = \Gamma_t \cup \Gamma_u$, $\Gamma_t \cap \Gamma_u = \emptyset$, $\mathbf{n}_\Gamma$ is the normal outward vector to the boundary (Figure 1). The boundary conditions for the fluid phase are written in terms of fluxes and pressures

$$n_f (\mathbf{v}_f - \mathbf{v}_s) \cdot \mathbf{n}_\Gamma = q_p \quad \text{en } \Gamma_q \quad p = p_p \quad \text{en } \Gamma_p \quad (2.5)$$

with $\Gamma = \Gamma_p \cup \Gamma_q$, $\Gamma_p \cap \Gamma_q = \emptyset$, and the initial conditions at $t = 0$ and $\pi = s, f$ are

$$\mathbf{u}_\pi (\mathbf{x}, 0) = \mathbf{u}_\pi^0 \quad p(\mathbf{x}, 0) = p^0 \quad (2.6)$$

In Figure 1 can be observed that the domain has a discontinuity $\Gamma_d$ where cohesive forces can be generated, such that:

$$\sigma \cdot \mathbf{n}_{\Gamma_d^+} = t^+ \quad \text{en } \Gamma_d^+ \quad \sigma \cdot \mathbf{n}_{\Gamma_d^-} = t^- \quad \text{en } \Gamma_d^- \quad (2.7)$$

where $t^+$ and $t^-$ are the cohesive forces acting over the opposite faces of the fracture $\Gamma_d^+$ and $\Gamma_d^-$ respectively. Since $\mathbf{n}_{\Gamma_d^+} = -\mathbf{n}_{\Gamma_d^-}$, there holds

$$\sigma \cdot \mathbf{n}^+ = -\sigma \cdot \mathbf{n}^- = t^+ = t^- = t_c - p \cdot \mathbf{n}_{\Gamma_d} \quad \text{en } \Gamma_d \quad (2.8)$$
2.2 Constitutive laws

In saturated porous media mechanics the following equations are used

\[ \dot{\sigma} = D : \dot{\varepsilon} - \alpha I \dot{\rho} \quad \dot{\phi} = \alpha I \dot{\varepsilon} - \frac{1}{K_f} \dot{\rho} \]  

(2.9)

where \( \sigma \) is the effective stress tensor, \( D \) is the elasticity tensor, \( \varepsilon = (\nabla u^T + \nabla u)/2 \) is the deformation tensor, \( \alpha \) is the Biot constant, \( I \) is the identity tensor, \( p \) is the pore pressure, \( \phi \) is the porosity and \( K_f \) is the Biot modulus. For the description of the flux through the porous media, the Darcy's law is adopted

\[ n_f (v_f - v_s) = -k_f \nabla p \]  

(2.10)

where \( k_f \) is the permeability of the porous media, \( n_f \) is the fraction of the fluid volume with respect to the total volume and \( \nabla p \) is the pressure gradient.

2.3 Weak form of the equation

In order to obtain the weak form of the balance equation, the momentum balance (2.1) and the mass balance (2.3) are multiplied by trial functions which are cinematically admissible with the displacements field, \( \delta u \), and with the pressures field, \( \delta p \), as follows

\[ \int_{\Omega} (\nabla \cdot \sigma) \delta u \, dv = 0 \]

\[ \int_{\Omega} \alpha \nabla \cdot \nu_s \delta p \, dv + \int_{\Omega} n_f \nabla \cdot (v_f - v_s) \delta p \, dv + \int_{\Omega} \frac{1}{K_f} \frac{\partial p}{\partial t} \delta p \, dv = 0 \]  

(2.11)

Using the divergence theorem and the boundary conditions (2.4), (2.5) and (2.7), the equations (2.11)_1 and (2.11)_2 can be written as follows
\[ \begin{align*}
\int_{\Omega} \sigma : \nabla(\delta u) \, dv - \int_{\Gamma_t} t_p \cdot \delta u \, da + \int_{\Gamma_d} (t_\text{c} - p n_{\Gamma_d}) [\delta u] \, da &= 0 \\
\int_{\Omega} \alpha \nabla \cdot v_s \delta p \, dv + \int_{\Omega} \frac{1}{K_f} \frac{\partial p}{\partial t} \delta p \, dv - \int_{\Omega} \nabla \delta p \cdot v_r \, dv + \int_{\Gamma_d} v_r \cdot n_d \delta p \, da + \int_{\Gamma_d} [v_r] \cdot n_{\Gamma_d} \delta p \, da &= 0
\end{align*} \] (2.12)

where \([\delta u] = (\delta u^- - \delta u^+)\) is the jump of the displacement field at the opposite faces of the crack and \(v_r = n_f(v_f - v_s)\) is the relative velocity between the fluid and the solid matrix.

The last term of the equation (2.12) represents the jump of the flux that goes through the faces of the fracture or discontinuity, which can be rewritten as \(\int_{\Gamma} [v_r \delta p] \cdot n_{\Gamma_d} \delta p \, da\), since the pressure is supposed constant over the crack width. In this work the cohesive forces are supposed null, which means \(t_\text{c} = 0\), since only the fluid pressure applied on both sides of the discontinuity will be taken into consideration.

### 2.4 Mass balance in the discontinuity

For the integration of the flux inside the fracture, the mass balance in the domain \(\Omega^*\) complementary to the domain \(\Omega\) is used (Figure 1). Supposing that in \(\Omega^*\), the Biot’s coefficient has a value of \(\alpha = 1\) and that \(n_f = 1\), which means that inside the fracture the fluid occupies all the volume. Thus, the mass balance in \(\Omega^*\) is

\[ \int_{\Omega^*} \nabla \cdot v_s \delta p \, dv + \int_{\Omega^*} (v_f - v_s) \delta p \, dv + \int_{\Omega^*} \delta p \frac{1}{K_f} \frac{\partial p}{\partial t} = 0 \] (2.13)

Using the divergence theorem and the Darcy’s law (2.10) the following equation is obtained

\[ \int_{\Gamma} [v_f - v_s] \cdot n^\Gamma \delta p \, da = - \int_{\Omega^*} \nabla \cdot v_s \delta p \, dv + \int_{\Omega^*} \nabla (k_d p) \cdot \nabla \delta p \, dv - \int_{\Omega^*} \frac{1}{K_f} \frac{\partial p}{\partial t} \delta p \, dv \] (2.14)

Considering that the crack opening is much smaller than the length of it and supposing that the tangential component of the velocity \(v_{sx^*}\) varies linearly with the \(y^*\) direction, the first integral of (2.13) over \(\Omega^*\) is calculated as follows

\[ \int_{\Omega^*} \nabla \cdot v_s \delta p \, dv = \int_{\Gamma_d} \int_{-h}^{h} \left( \frac{\partial v_{sx^*}}{\partial x^*} + \frac{\partial v_{sy^*}}{\partial y^*} \right) \delta p \, dy^* \, da \]

\[ = \int_{\Gamma_d} \left( 2h \left( \frac{\partial v_{sx^*}}{\partial x^*} \right) + [v_y^*] \right) \delta p \, da \] (2.15)

where the symbol \(\langle \cdot \rangle\) represents the average value of the function. Similarly, the tangential derivative of the pressure does not vary with the \(y^*\) direction and the normal derivative disappears because the pressure is considered constant in the fracture high. Therefore, the second integral of (2.13) is written
\[ \int_{\Omega} \nabla (k_d p) \cdot \nabla \delta p \, dv = \int_{\Gamma_d} \int_{-h}^{h} k_d \nabla p \cdot \nabla \delta p \, dy \, da = \int_{\Gamma_d} 2h k_d \frac{\partial p}{\partial x^*} \frac{\partial \delta p}{\partial x^*} \, da \] (2.16)

Similarly, the third integral is solved as
\[ \int_{\Gamma_d} \frac{1}{Q} \frac{\partial p}{\partial t} \delta p \, dv = \int_{\Gamma_d} 2h \frac{1}{Q} \frac{\partial p}{\partial t} \delta p \, da \] (2.17)

and the flux inside the crack is written as
\[ \int_{\Gamma_d} [v_f - v_s] \cdot n \, \delta p \, da = -\int_{\Gamma_d} 2h \left( \frac{\partial v_{x*}}{\partial x^*} \right) \delta p \, da - \int_{\Gamma_d} \left[ v_{sy*} \right] \delta p \, da + \int_{\Gamma_d} 2h k_d \frac{\partial p}{\partial x^*} \frac{\partial \delta p}{\partial x^*} \, da - \int_{\Gamma_d} 2h \frac{1}{K_f} \frac{\partial p}{\partial t} \delta p \, da \] (2.18)

### 3 NUMERICAL SOLUTION OF THE HYDROMECHANICAL COUPLED PROBLEM

For the solution of the problem, the eXtended Finite Element Method (XFEM) will be presented in this section. The XFEM has been developed with the aim of taking into account discontinuities, and its propagation, in the displacement field due to fractures, for example by enlarging the space of the displacement field through the introduction of enrichment functions. This method has been extended to the case of fluid transport by enriching also the pressure field.

#### 3.1 Modelling voids, strong and weak discontinuities in XFEM

XFEM describes a field as the sum of a linear combination of functions of regular forms and other linear combination of enrichment functions. This method allows to generate fields with presence of voids, strong and weak discontinuities using meshes unadjusted to the geometry of these irregularities. The choice of the functions of the enrichment forms depends on the discontinuity type to be reproduced. A strong discontinuity presents a jump in the field under study, as in the case of a discrete fracture. In order to generate that discontinuous field, the functions of enrichment form have to be discontinuous, for example the Heaviside’s function (3.1). In a weak discontinuity the field under study is continuous but the field normal gradient is discontinuous. The Distance function (Heaviside function) (3.1), which is continuous with discontinuous derivative, satisfies these requirements. The element that contains the crack-tip requires special enrichment functions that fulfill the same requirements. For this, the bases of the analytical solutions of each problem under study are used and the equation (3.2) and (3.2) are obtained for strong and weak discontinuities respectively. Finally, for voids consideration, a Heaviside-type enrichment function is used, where a node that lies outside the void will have \( V(X) = 1 \) and a node that lies inside the void will have \( V(X) = 0 \). The nodes that lies inside the void and
whose nodal support is not intersected by the void are removed from the calculations.

\[ H = \begin{cases} 1 & \text{if } x \in \Omega^+; \\ -1 & \text{if } x \in \Omega^-; \end{cases} \]

\[ Z = \begin{cases} d & \text{if } x \in \Omega^+; \\ -d & \text{if } x \in \Omega^-; \end{cases} \]

\[ \{F(r, \theta)\}_{r=1}^3 = \{\sqrt{r} \cos \theta/2, \sqrt{r} \sin \theta/2, \sqrt{r} \sin \theta/2 \sin \theta, \sqrt{r} \cos \theta/2 \sin \theta\} \]

\[ G(r, \theta) = \sqrt{r} \sin \theta/2 \]

### 3.2 Space and time discretization of the coupled problem

As to spatial discretization, the domain \( \Omega \) is discretized into finite elements, where for each specific element the displacement and pore pressure are interpolated by using shape functions \( N_{ar}^I \) and \( N_{pr}^I \) (with \( I \) the node within the element and \( r = 1 \) for standard interpolation functions and \( r = 2, 3 \) for enrichment function associated with discontinuity and near tip singularity interpolation functions, respectively), of the form

\[
\begin{align*}
\mathbf{u} &\approx \mathbf{u}^h(x) = \sum_{i=1}^I N_{a1}^i(x) \mathbf{u}_1 + \sum_{j \in J} N_{a2}^j(x) \mathbf{u}_2 + \sum_{k \in K} N_{a3}^k(x) \mathbf{u}_3 \\
p &\approx p^h(x) = \sum_{i=1}^I N_{p1}^i(x) p_1 + \sum_{j \in J} N_{p2}^j(x) p_2 + \sum_{k \in K} N_{p3}^k(x) p_3
\end{align*}
\]

where \( I \) denote the set of standard finite element nodes, \( J \) enrichment nodes related to the Heaviside step function and \( K \) enrichment nodes related to the near tip asymptotic function, \((\mathbf{u}_1^i, p_1^i)\) conventional displacement and pressure values of node \( i \), \((\mathbf{u}_2^i, p_2^i)\) and \((\mathbf{u}_3^i, p_3^i)\) are the additional nodal values of displacement and pressure related to the following shape functions

\[
\begin{align*}
N_{a2}^j(x) &= N_{a1}^j (H(x) - H(x_j)) \\
N_{a3}^k(x) &= N_{a1}^k (F(x) - F(x_k)) \\
N_{p2}^j(x) &= N_{p1}^j (Z(x) - Z(x_j)) \\
N_{p3}^k(x) &= N_{p1}^k (G(x) - G(x_k))
\end{align*}
\]

where \( H \) denote the Heaviside step shape function centered at the discontinuity and \( Z \) denote the distance functions, \( F \) and \( G \) are enhanced basis defined in previous sections.

In order to accommodate the propagation of discrete cracks through elements, Belytschko used the partition of unity property of the finite element shape function \( N \), i.e. \( \sum_{i=1}^n N_{ai} = 1 \) with \( n \) the number of nodal points in an element. The crack is then seen as a discontinuity in the displacement field, and the displacement can be written as the sum of two continuous displacement fields

\[
\begin{align*}
\mathbf{u} &= N_{a1} \mathbf{a}_1 + N_{a2} \mathbf{a}_2 + N_{a3} \mathbf{a}_3 \\
\nabla \mathbf{u} &= B_{a1} \mathbf{a}_1 + B_{a2} \mathbf{a}_2 + B_{a3} \mathbf{a}_3 \\
p &= N_{p1} p_1 + N_{p2} p_2 + N_{p3} p_3
\end{align*}
\]

where \( N \) and \( B = \nabla N \) denotes the standard shape functions and gradient of the shape functions respectively, whereas \( \mathbf{a}_i \), \( i = 1, 2, 3 \), contains the conventional and additional degrees of freedom.
\[ v_s = N_{a_1} \dot{a}_1 + N_{a_2} \dot{a}_2 + N_{a_3} \dot{a}_3 \nabla \cdot v_s = m^T (B_{a_1} a_1 + B_{a_2} a_2 + B_{a_3} a_3) \]
\[ \dot{p} = N_{p_1} \dot{p}_1 + N_{p_2} \dot{p}_2 + N_{p_3} \dot{p}_3 \nabla \cdot \dot{p} = B_{p_1} \dot{p}_1 + B_{p_2} \dot{p}_2 + B_{p_3} \dot{p}_3 \]

\[
\int_{\Omega} B_{a_1}^T \sigma \, dV + \int_{\Gamma} p N_{a_1}^T n_\Gamma \, ds - \int_{\partial \Omega} N_{a_1}^T \dot{t} \, ds = 0
\]
\[
\int_{\Omega} N_{p_1}^T \nabla \cdot v_s \, dV + \int_{\Omega} N_{p_1}^T \frac{\partial p}{\partial t} \, dV - \int_{\Omega} B_{p_1}^T w \, dV + \int_{\Gamma} N_{p_1}^T [w] \cdot n \, ds + \int_{\partial \Omega} N_{p_1}^T q \, ds = 0
\]

(3.6)

with \( i = 1, 2, 3 \). Traction, pressure and pressure test function have unique values across the discontinuity.

The constitutive laws are then introduced.

\[ \sigma = D : \varepsilon - \alpha p 1 \quad w = -k \nabla p \]  

(3.8)

The following coupled set of equations can be obtained as

\[ C \dot{x} + K x + f = 0 \]  

(3.9)

where \( x^T = [a_1, a_2, a_3, p_1, p_2, p_3] \) vector of nodal unknowns.

The time discretization using the backward Euler integrated format of the mass and linear momentum balance relation has the form

\[ F_{n+1} = C_{n+1} \frac{\Delta x}{\Delta t} + K_{n+1} x_{n+1} + f_{n+1} = 0 \]  

(3.10)

with \( \Delta x = x^{n+1} - x^n \), \( \Delta t = t^{n+1} - t^n \), and \( (x^{n+1}, x^n) \) denote the unknowns at time \( t^{n+1} \) and \( t^n \), respectively. The equations as stated in (3.10) are conveniently solved using Newton–Raphson iterations.

### 3.3 Crack initiation and growth

The accuracy and reliability of the analysis of a cracked body primarily depends upon the accurate determination and continuity of the crack path. The approach used in this work is that the Mode I Stress Intensity Factor reaches the critical stress \( K_I \) intensity factor \( K_{IC} \), obtained from the classical linear fracture mechanics by \( K_{IC} = \sqrt{2E} \), where \( E \) is the elasticity modulus. Let us assume that the imposed actions \( Q \) depend linearly on a scalar parameter (load factor) \( \lambda \) such that \( Q = \lambda Q_0 \). The load factor is found such that \( K_I \) reaches \( K_{IC} \). To determine the direction of propagation, the criterion is adopted maximum hoop stress criterion \( [5] \), which defines the propagation direction to be along the normal direction to maximum hoop stress, resulting

\[ \theta = 2 \arctan \frac{1}{4} \left( \frac{K_I}{K_{II}} \pm \sqrt{\left( \frac{K_I}{K_{II}} \right)^2 + 8} \right) \]  

(3.11)
where \( K_I, K_{II} \) are the Mode I and Mode II stress intensity factor respectively.

4 \hspace{1mm} \text{NUMERICAL EXAMPLES}

4.1 \hspace{1mm} \text{Mechanical fracture propagation}

In the first example, an elastic linear media with a Young’s modulus \( E = 30 \) and a Poisson’s coefficient \( \mu = 0.2 \), with lengths \( L_x = 2000 \) and \( L_y = 2000 \) and a circular void with a diameter \( D = 1000 \) whose center is displaced is considered (Figure 2.a). The boundary conditions are: the restriction of the lower boundary displacement and a mechanical variable load is imposed at the upper face \( \lambda \sigma \). The domain of the XFEM-based model has been discretized using 908 elements. The global response of the structure is shown in Figure 2.b in terms of the applied stress versus the opening displacement (vertical enlargement of the hole). In this figure, the segment (1) \( - \) (2) shows and increase of the load that produces stress concentration over the points \( A \) and \( B \); since the hole is not centered, larger stresses are obtained in the point \( B \); the point (2) shows the time in which the point \( B \) reaches the fail condition (according to \( K_{1c} \)) and goes on propagating to the free boundary which leads to the segment (2) \( - \) (3). Then, the left crack starts opening again (segment (3) \( - \) (4)) before finally propagating (segment (4) \( - \) (5)).

4.2 \hspace{1mm} \text{Hydraulic fracture propagation}

In the second example, a linear elastic media with the same characteristics of the first example is considered, but in this case the mechanical behaviour is coupled to the fluid flow

\[ \text{Figure 2: (a) First example scheme. (b) Global response of the specimen: } \lambda \sigma \text{ vs crack opening displacement } w_h. \]
inside the porous matrix. The specimen is made of porous fluid saturated material with a permeability $k_m = 1$, Biot’s coefficient $\alpha = 1$, compressibility module $K_f = 1E18 GPa$ and dimensions $L_x = 2000$ and $L_y = 2000$ (Figure 4.a). The boundary conditions are: the restriction of the lower boundary displacement -in order to remove rigid body motions- and null boundary pressure. A fluid $\lambda Q$ is imposed at a preexisting crack with a length $a = 400$, as shown in figure 4.a. The model has been discretized using 420 mixed finite elements with biquadratic-shaped functions for the displacement field and bilinear ones for the pressure field. The global response of the structure is given in Figure 4.b in terms of the applied pressure versus the crack opening displacement. The segment (1) – (2) shows a pressure increase produced by an increase of the fluid injection, which leads, due to the coupling with the displacement field, to stresses in the crack tip that increase the stress intensity factor up to a critic value -point (2)-. From this point, the crack goes on propagating to the hole. In figure 4.b the pressures distribution along a line parallel to the initial crack for different positions of the crack tip are shown.

5 CONCLUDING REMARKS

In this paper the problem of fluid flow through a fractured deformable porous media has been formulated. With the aim of solving numerically this problem in a coupled way, XFEM method has been choose. The hydro-mechanical coupling of the porous media has been modelled using the Biot-Coussy theory. A formulation that uses an extension of the Poiseuille’s law for viscous fluids to the flow between two parallel plates has been developed. For the crack propagation analysis, the stress intensity factor criterion from LEFM has been assumed. Then, in order to verify the theoretical findings, two examples have been developed. In the first, a crack propagation due to a mechanical load in a
Figure 4: (a) Second example scheme. (b) Media deformation produced by fluid injection when the crack-tip is \( y = 400 \).

Figure 5: (a) Global response of the specimen: Fluid injection \( \lambda Q \) vs crack opening displacement \( w_h \). (b) Pressure along a line parallel to the crack for different positions of the crack-tip.
homogenous media with a hole where two cracks went on propagating has been simulated. The load vs COD curve and the deformed models in different simulation times have been shown. In the second example, a crack propagation due to a fluid injection through an preexisting crack has been simulated. The injection fluid vs COD curve and the pressures distribution for different positions of the crack-tip have been shown. The obtained results have been satisfactory.

REFERENCES


GLOBAL METHOD OF APPROXIMATE PARTICULAR SOLUTIONS FOR SOLVING THE RECTANGULAR TWO-SIDED LID-DRIVEN CAVITY PROBLEM FOR TWO-DIMENSIONAL AND INCOMPRESSIBLE FLOW

J.M. GRANADOS1, C.A. BUSTAMANTE2, H. POWER3, W.F. FLOREZ2 AND C. NIETO2

1Instituci´on Universitaria de Envigado
055420 Envigado, Colombia
E-mail: jmgranados@correo.iue.edu.co, Web page: http://www.iue.edu.co

2School of Engineering
Universidad Pontificia Bolivariana
Campus Laureles, 050031 Medellin, Colombia
Email: carlos.bustamante@upb.edu.co; whady.florez@upb.edu.co; cesar.nieto@upb.edu.co
Web page: http://www.upb.edu.co

3School of Mechanical, Materials and Manufacturing Engineering
University of Nottingham
University Park, Nottingham NG7 2RD, UK
Email: henry.power@nottingham.ac.uk - Web page: http://www.nottingham.ac.uk

Key words: Particular solutions, Meshless method, RBF

Abstract. The two-dimensional Navier-Stokes equations in primitive formulation are solved by the global Method of Approximated Particular Solutions (MAPS). The proposed scheme provides a good treatment of the nonlinearities in the momentum equations by using the Newton-Raphson method with variable step. Radial Basis Functions (RBFs) are used as source terms in the Stokes equations in order to obtain particular solutions for the linear part of the momentum equations. A closed-form expression of the particular solutions as a function of domain coordinates are obtained by employing Oseen’s decomposition formula. A variation of the classical lid-driven cavity problem, the rectangular two-sided lid-driven cavity flow, is solved with the numerical scheme developed. In this case, the movement of fluid occurs as a result of the constant movement of the two vertical sides. The flow pattern is related to the cavity aspect ratio and the Reynolds number, Re. In this paper, a rectangular cavity with constant height equal to 1 and Re = 700 are considered and the variation in cavity width, 1 ≤ l ≤ 2.5, is analysed. Numerical results reproduce the flow pattern change when increasing l = 1.995 to l = 2 showing good agreement with results reported in the literature.
1 INTRODUCTION

Classical numerical methods such as Finite Difference (FDM), Finite Element (FEM) and Finite Volume (FVM) have been widely used to solve fluid flows. However, the time required to generate a mesh of good quality or to remesh [1] and the difficulties associated to the dimension of the problem, have caused that the meshless methods have become a suitable alternative. In particular, meshless methods using Radial Basis Functions (RBFs) have been studied in recent years because they are not only easy to extend to three-dimensional problems [2, 3] but also easy to implement, have high convergence rates and flexibility with regards to the enforcing of boundary conditions.

The RBFs was introduced by Hardy in [4] to approximate two-dimensional geographical surfaces from a scattered data set. Franke [5] evaluated 29 interpolation algorithms based on accuracy, sensitivity to parameters, storage requirements and ease of implementation. This analysis found that global methods outperformed local methods and among the methods tested, the Hardy’s Multiquadric (MQ) RBFs method obtained the most accurate results. Kansa [6] used MQ RBFs to develop a global meshless method for the numerical solution of Partial Differential Equations (PDEs) in which the unknown variable is approximated directly with RBFs interpolations. This idea was subsequently used to solve other PDEs, such as Poisson, [7], and Navier-Stokes, [8].

Mai-Duy and Tran-Cong [11, 12] proposed to use RBFs to approximate the higher order derivatives of the PDE, instead of the dependent variables of the problem, as in the method of Kansa. By direct integration of the derivatives approach the variables of the problem are obtained. This scheme, called indirect approach, shows greater accuracy and better stability, in terms of shape parameter, than the direct approach due to the smooth behaviour of the integration process, which, unlike the derivation, does not contain inherent inaccuracy of the approximation.

Chen et al. [13] proposed an integrated RBF method for the case of a linear PDE, in which a differential operator, or only part of it, is given in terms of the radial component of a polar or spherical coordinate systems. In this case, these authors proposed to approximate the radial component of the PDE in terms of an RBF interpolation, and, by integrating the resulting non-homogeneous ordinary differential equation, it is possible to obtain an approximated representation of the field variable by a linear superposition of the corresponding particular solution. Superposition of particular solutions has been used previously in connection with the numerical solution of linear boundary value problems, using the classical decomposition form in terms of its particular and homogeneous solutions, combined with the method of fundamental solutions (see [14]) and with the boundary element method (see [15]).

Recently, Bustamante et al. in [16] developed the Method of Approximated Particular Solutions (MAPS) to solve the Navier-Stokes system of equations in terms of its primitive variable formulation, where the velocity and pressure fields are defined in terms of superposition of Stokes particular solutions, with a MQ RBF as nonhomogeneous term,
obtained from the Oseen’s decomposition formula [17], and the use of a direct iterative scheme (Picard iterations) to deal with corresponding non-linear terms of the governing equation. By using the proposed global MAPS, accurate results were obtained for some steady-state benchmark problems with incompressible flow at high \( Re \). In this approach, it is not necessary to define any pressure-velocity coupling strategy since the pressure particular solution is directly found from the obtained expression of the velocity particular solution, besides in the construction of the global collocation matrix system it is not necessary to impose the continuity equation since velocity particular solution exactly satisfies the equation of continuity.

In this paper, the method developed in [16] is improved by using the Newton-Raphson method and its applicability is widened by implementing temporal discretization. The Taylor’s decaying vortices transient problem was used to validate the proposed method and the two-sided lid-driven cavity transient problem is solved for rectangular cavities at \( Re = 700 \).

2 METHOD OF APPROXIMATED PARTICULAR SOLUTIONS (MAPS)

Let \( \Omega \) be an arbitrary two-dimensional region and \( \Gamma \) its boundary. The two-dimensional Navier-Stokes equations for incompressible flow consists of the continuity equation and the momentum equations, which are given by:

\[
\frac{\partial u_i(\vec{x})}{\partial x_i} = 0, \tag{1}
\]

\[
\rho \frac{\partial u_i(\vec{x})}{\partial t} = \mu \frac{\partial^2 u_i(\vec{x})}{\partial x_j \partial x_j} - \rho \frac{\partial p(\vec{x})}{\partial x_i} - \rho u_j(\vec{x}) \frac{\partial u_i(\vec{x})}{\partial x_j}, \tag{2}
\]

where \( \vec{x} = (x_1, x_2) \in \Omega, i = 1, 2, \) and \( \rho, p \) and \( \vec{u} = (u_1, u_2) \) are, respectively, density, pressure and velocity field. The boundary conditions for \( \vec{x} \in \Gamma \) are defined in terms of given values of the velocity and/or surface traction, i.e.,

\[
u_i(\vec{x}) = u_{ib}, \tag{3}
\]

\[
\tau_i(\vec{x}) = \sigma_{ij}(\vec{x}) n_j = -p(\vec{x}) n_i + \mu \left( \frac{\partial u_i(\vec{x})}{\partial x_j} + \frac{\partial u_j(\vec{x})}{\partial x_i} \right) n_j = \tau_{ib}. \tag{4}
\]

The MAPS approximates the velocity and pressure fields as a linear combination of a set of functions \( \hat{u}_i^l \) and \( \hat{p}^l \), with \( i, l = 1, 2 \),

\[
u_i(\vec{x}) = \sum_{k=1}^{N} a_k^l \hat{u}_i^l(\vec{x}, \vec{\xi}_k) \quad \text{and} \quad p(\vec{x}) = \sum_{k=1}^{N} a_k^l \hat{p}^l(\vec{x}, \vec{\xi}_k), \tag{5}
\]

which depends on the Euclidean distance, \( r_k \), between \( \vec{x} \) and \( \vec{\xi}_k \), with \( \vec{\xi}_k \) as the \( k \)-th test point. The particular velocities and pressures, \( \hat{u}_i^l \) and \( \hat{p}^l \), are particular solutions of the Stokes problem (6)-(7) where the non-homogeneous term is defined as the RBF
\[ \phi(r) = \sqrt{r^2 + c^2}. \] In order to obtain a closed expression for the particular solutions, we use the Oseen’s decomposition formula (see [17]) to solve the Stokes equations system (6)-(7).

\[ \frac{\partial \tilde{u}_i^l(x, \xi)}{\partial x_i} = 0, \tag{6} \]

\[ \frac{\partial^2 \tilde{u}_i^l(x, \xi)}{\partial x_j \partial x_j} - \frac{\partial \tilde{p}_l(x, \xi)}{\partial x_i} = \phi(r) \delta_{il}. \tag{7} \]

First of all, we define the particular velocities in terms of an auxiliary potential \( \psi \),

\[ \tilde{u}_i^l(x, \xi) = \frac{\partial^2 \psi(r)}{\partial x_m \partial x_m} \delta_{il} - \frac{\partial^2 \psi(r)}{\partial x_i \partial x_l}. \tag{8} \]

By substituting (8) into (7), expressions for the auxiliary potential in terms of the RBF (9) and for the particular pressures as a function of the auxiliary potential (10) are obtained. Due to the definition of the particular velocity field (equation (8)), the continuity equation is satisfied automatically.

\[ \mu \frac{\partial^4 \psi(r)}{\partial x_m \partial x_m \partial x_k \partial x_k} = \phi(r) \tag{9} \]

\[ \mu \frac{\partial^4 \psi(r)}{\partial x_m \partial x_m \partial x_i \partial x_l} + \frac{\partial \tilde{p}_l(x, \xi)}{\partial x_i} = 0. \tag{10} \]

By direct integration of the first of these two equations, the following expression for \( \psi \) in terms of \( r \) is obtained.

\[ \frac{\partial \psi(r)}{\partial x_m \partial x_m} \delta_{il} - \frac{\partial \psi(r)}{\partial x_i \partial x_l}. \]
\[ \psi(r) = \frac{1}{12\mu} \left[ \frac{1}{75} \phi(r) \left( 4r^4 + 48r^2c^2 - 61c^4 \right) - c^3 \ln(c) r^2 - \frac{1}{5} \left( 5r^2 - 2c^2 \right) c^3 \ln[c + \phi(r)] \right]. \]  

(11)

By substituting (11) into (8) and into (10), expressions for \( \hat{u}_l \) and \( \hat{p}_l \) in terms of \( r \) are achieved for being used in (5) to approximate \( u_i \) and \( p \) (see [16]).

Replacing expressions (5) into steady momentum equations leads to the following nonlinear equation in terms of the particular solutions

\[
\sum_{k=1}^{N} \alpha_{lk} \left[ \mu \frac{\partial^2 \hat{u}_l(r_k)}{\partial x_j \partial x_i} - \frac{\partial \hat{p}_l(r_k)}{\partial x_i} \right] - \rho \sum_{k=1}^{N} \alpha_{lk} \hat{u}_j(r_k) \sum_{k=1}^{N} \alpha_{lk} \frac{\partial \hat{u}_l(r_k)}{\partial x_j} = 0. \]  

(12)

From direct substitution of (7) in the previous equation and by using matrix notation, the following momentum equation is attained

\[
\begin{bmatrix}
[\phi] \\
[0]
\end{bmatrix}
\begin{bmatrix}
[\alpha_1^1] \\
[\alpha_2^1]
\end{bmatrix}
- \rho \begin{bmatrix}
[\alpha_1^1] \\
[\alpha_2^1]
\end{bmatrix}
\begin{bmatrix}
\partial \hat{u}_1 \\
\partial x_j
\end{bmatrix}
\begin{bmatrix}
[\alpha_1^1] \\
[\alpha_2^1]
\end{bmatrix}
= 0, \]  

(13)

where

\[
\begin{bmatrix}
\hat{u}_j \\
\frac{\partial \hat{u}_i}{\partial x_j}
\end{bmatrix}
= \begin{bmatrix}
\hat{u}_1^1(r_1), \ldots, \hat{u}_1^1(r_N), \hat{u}_j^2(r_1), \ldots, \hat{u}_j^2(r_N)
\end{bmatrix},
\]  

(14)

\[
\begin{bmatrix}
\frac{\partial \hat{u}_1}{\partial x_j} \\
\frac{\partial \hat{u}_2}{\partial x_j} \\
\frac{\partial \hat{u}_3}{\partial x_j} \\
\vdots \\
\frac{\partial \hat{u}_N}{\partial x_j}
\end{bmatrix}
= \begin{bmatrix}
\frac{\partial \hat{u}_1^1(r_1)}{\partial x_j}, \ldots, \frac{\partial \hat{u}_1^1(r_N)}{\partial x_j}, \frac{\partial \hat{u}_2^1(r_1)}{\partial x_j}, \ldots, \frac{\partial \hat{u}_2^1(r_N)}{\partial x_j}
\end{bmatrix}.
\]  

The unknown variables in (13) are the coefficients, \( \alpha_{lk}^i \), of the approximations (5). The coefficients have no physical sense, and in consequence there is no criteria for choosing

\[ Figure 2: \] RMS error at \( t = 1.5 \) for \( 5 \times 5, 15 \times 15, 26 \times 26 \) and \( 35 \times 35 \) uniform nodal distributions.
the values of the coefficients for the initial guess in the Newton’s method. That is reason of using a direct formulation whereby the unknowns are the values of the velocity field. As it is customary in this type of global meshless schemes with RBFs, the collocation nodes are the same as the test points with \( N_b \) boundary nodes, \( N_i \) internal nodes and \( N_b + N_i = N \). The expression (15), for the coefficients of the interpolation (5) in terms of the 2\( N \) unknown values of the velocity field over the region \( \Omega \), is obtained from the collocation of the proposed interpolations for \( u_i, i = 1, 2 \) at each of the nodes in the region.

\[
\begin{bmatrix}
\hat{u}_1 \\
\hat{u}_2 \\
\alpha_1 \\
\alpha_2
\end{bmatrix}
= \begin{bmatrix}
u_1 \\
u_2 \\
1 \\
1
\end{bmatrix} \iff \begin{bmatrix}
\alpha_1 \\
\alpha_2
\end{bmatrix} = \hat{u}^{-1} \begin{bmatrix}
u_1 \\
u_2 \\
1 \\
1
\end{bmatrix}. \tag{15}
\]

If (15) is substituted in (13) and the notation introduced in (14) is used, the convective term in the momentum equation is simplified as shown in the following expression which is the momentum equation in terms of the velocity field

\[
[\phi \delta u] [\hat{u}]^{-1} \begin{bmatrix} u_1 \\
u_2 \end{bmatrix} - \rho u_j \frac{\partial \hat{u}_i}{\partial x_j} [\hat{u}]^{-1} \begin{bmatrix} u_1 \\
u_2 \end{bmatrix} = 0. \tag{16}
\]

The set of 2\( N \) functional relationships required by Newton-Raphson method is obtained after collocation process of expressions (3) and/or (4) on the \( N_b \) boundary nodes and equation (16) on the \( N_i \) internal nodes. The functional relationship vector, \( [F] = ([F^1] [F^2])^T \), is constructed considering that

\[
F_k^i \left( \begin{bmatrix}
u_1 \\
u_2 \end{bmatrix} \right) = \begin{cases}
u_i - \nu_{ib} & \text{si } \bar{x}^k \in \Gamma_u, \\
\tau_i - \tau_{ib} & \text{si } \bar{x}^k \in \Gamma_\tau, \\
[\phi \delta u] [\hat{u}]^{-1} \begin{bmatrix} u_1 \\
u_2 \end{bmatrix} - \rho u_j \frac{\partial \hat{u}_i}{\partial x_j} [\hat{u}]^{-1} \begin{bmatrix} u_1 \\
u_2 \end{bmatrix} & \text{si } \bar{x}^k \in \Omega,
\end{cases}
\]

where \( \bar{x}^k \) is the \( k \)-th node with \( 1 \leq k \leq N \). We denote with \( \Gamma_u \) and \( \Gamma_\tau \) the sets of nodes at which velocity condition or surface traction condition are imposed, respectively. According to the Newton-Raphson method, \( \hat{F} \) allows to obtain a new approach, \( [w_{ne}] = ([u_1] [u_2])^T \), from a known approach, \( [w_{kn}] \), through the relation \( \delta \bar{w} = \bar{w}_{ne} - \bar{w}_{kn} + \delta \bar{w} \) where \( \delta \bar{w} \) is the step’s correction which is given by equation (17) with the goal of moving each component of \( \hat{F} \) closer to zero.

\[
[J] \cdot [\delta \bar{w}] = -[\hat{F}] \tag{17}
\]

In order to reduce the computation time, the arrays \( [\phi \delta u] [\hat{u}]^{-1} \) and \( \rho \left[ \frac{\partial \hat{u}_i}{\partial x_j} \right] [\hat{u}]^{-1} \) are constructed and stored before starting the Newton-Raphson iterations. The consumed time at each iteration, spent almost entirely to build the Jacobian matrix, can be decreased if parallelization strategies are used to construct the columns of the Jacobian matrix more efficiently.
3 NUMERICAL RESULTS

In this section, we use the MAPS to solve the two-dimensional Navier-Stokes equations (1)-(2) for incompressible flow according to the algorithm presented above. Two transient problems, one of them with analytical solution, are solved. In previous papers, (see [16], [19]), the behaviour of the method in terms of the shape parameter was evaluated and the obtained results indicated that the global MAPS is relatively insensitive to changes of the value, within a wide range of values between $1 \times 10^{-5}$ and $1 \times 10^{-1}$.

Solutions for different values of the shape parameter were obtained and better results for the global MAPS with Newton-Raphson method are observed at $1 \times 10^{-5}$. In all problems solved, a shape parameter $c = 1 \times 10^{-5}$ is used and the Reynolds number is defined as $Re = \rho U L / \mu$, where $\rho$, $\mu$ and $\mu/\rho$ are the density, the dynamic viscosity and the kinematic viscosity of the fluid. $U$ and $L$ are the characteristics velocity and length, respectively. The physical conditions imposed in each case lead to the constant values $U = L = \mu = 1$ whereupon the $Re$ value is given by the fluid density value.

![Figure 3: Nodal distributions in the rectangular cavity at $Re = 700$ with a) $L = 1.955$ and b) $L = 2$](image)

3.1 TAYLOR’S DECAYING VORTICES

The Taylor’s decaying vortex problem has the analytical solution given by expressions (18) at $Re = 1$. The boundary conditions are Dirichlet type and they are obtained by evaluating the expressions for $u_1$ and $u_2$ on boundaries of a specified domain, in this case, the square $[0, 2\pi] \times [0, 2\pi]$.

In Figure (1) the velocity vectors for $t = 0$ (left) and the values of $u_1$ (right) on the vertical centre line of the square for different time steps are shown. Good agreement between numerical and analytical solutions is observed. The numerical solution was cal-
culated with a $16 \times 16$ uniform nodal distribution. Only the values obtained for $u_1$ are shown since $u_2$ has identical behaviour. Table 1 shows the RMS errors for $u_1$ at each time step considered in figure (1). Wang [20] reports an RMS error of $1.1 \times 10^{-3}$ for $u_1$ at $t = 0.4$ using a local scheme and the same nodal distribution to that used here, which indicates that excellent results have been obtained by the global MAPS. In addition, good agreement between numerical and analytical velocities are achieved in comparison to other author solutions obtained with finer nodal distributions. For instance, Shah [21] generated a $65 \times 65$ nodal distribution in such a way that maximum number of points get allocated to those regions with large velocity gradient. Due to asymptotic behaviour of the expressions (18) respect to $t$ errors decrease as time progresses.

$$
\begin{align*}
  u_1(x_1, x_2, t) &= -\cos(x_1)\sin(x_2)e^{-2t}, \\
  u_2(x_1, x_2, t) &= \sin(x_1)\cos(x_2)e^{-2t}, \\
  p(x_1, x_2, t) &= -\frac{1}{4} [\cos(2x_1) + \cos(2x_2)] e^{-4t}
\end{align*}
$$

The order of convergence of the spatial discretization scheme is analysed with the Taylor’s decaying vortices problem on the square $[0, \pi] \times [0, \pi]$, based on the conditions used by Botella [22]. With this aim, the RMS error for $t = 1.5$, with $5 \times 5$, $15 \times 15$, $26 \times 26$ and $35 \times 35$ nodal distributions, are calculated and shown in Figure (2). A potential regression of this data shows a convergence rate of the second order (equal to 2.1668). In all cases, a time step of $t = 0.01$ is used.

<table>
<thead>
<tr>
<th>$t$</th>
<th>$0.01$</th>
<th>$0.2$</th>
<th>$0.4$</th>
<th>$0.8$</th>
<th>$2.0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\epsilon_{RMS}$</td>
<td>$8.07 \times 10^{-4}$</td>
<td>$2.6576 \times 10^{-3}$</td>
<td>$2.43104 \times 10^{-3}$</td>
<td>$1.65424 \times 10^{-3}$</td>
<td>$4.45744 \times 10^{-4}$</td>
</tr>
</tbody>
</table>

**Table 1**: RMS errors between numerical and analytical solutions for $u_1$ at $Re = 1$

### 3.2 Rectangular Two-Sided Lid-Driven Cavity Flow

A variation of the lid-driven cavity problem obtained when two sides of the cavity are moved instead of just one. The cavity height is 1 and the cavity length, $l$, varies between 1 and 2.5. The two-sided lid-driven cavity flow is studied at $Re = 700$ with special attention in values of $l$ close to 2 in order to verify the change in the flow pattern at $l = 2$ [23]. Dirichlet boundary conditions are assumed for the velocity field. On the left and right vertical sides $u_2 = 1$ and $u_2 = -1$ while on the top and bottom sides $u_2 = 0$ and $u_1 = 0$.

Except for the rectangular cavity with $l = 2$, in all other cases, $51 \times 31$ refined towards-boundary nodal distributions are used. When $l = 2$, a $71 \times 41$ nodal distribution refined towards the boundaries and towards the vertical centreline of the cavity is used. Figures 3a) and 3b) show the nodal distributions used at $l = 1.955$ and $l = 2$.

A numerical study of the laminar, viscous and subsonic compressible flow in a two-dimensional, two-sided, lid-driven cavity using a multi-domain spectral element method
was presented by Shah et al. [23]. They investigated the effects of the cavity aspect ratio and the Mach number, $Ma$, on the flow. Their simulations showed that the streamline topologies at steady-state do not depend on $Ma$ for $Ma \leq 0.4$. Also, the same velocity field at steady-state is obtained for these low $Ma$ values if the flow is assumed incompressible. Our numerical scheme, is used to simulate incompressible flow in the two-sided lid-driven cavity in transient-state. The approximated streamlines at steady-state obtained with transient scheme are shown in Figure 4 which are in good agreement with those presented by [23]. When $l < 2$, a recirculation zone is produced in the cavity center at $Re = 700$. For $l \geq 2$, three recirculation zones are generated in the cavity. In each case, the centre lines of the cavity are highlighted with dotted lines for easy identification of the symmetry in the flow pattern.

![Figure 4: Approximated streamlines at steady state for $Re = 700$ at: a) $L = 1$, b) $L = 1.955$, c) $L = 2$, d) $L = 2.5$](image)

The Navier-Stokes system of equations (1) - (2) is solved with a zero velocity field as initial condition until the flow reaches a steady-state. Variable time steps are used in order to ensure the convergence of Newton’s method at each time step. Initially, small vortices are generated near the moving walls according to boundary conditions imposed, i.e., moving wall in contact with wall at rest at the corners of the cavity. Subsequently, the size of these vortices increases and the vortex configurations, shown in figure 4, are
attained. Evolution pattern of flow for $l = 1.955$, similar to that reported by Shah, is obtained and shown in Figure 5.

![Flow evolution at $Re = 700$ with $L = 1.955$](image)

**Figure 5**: Flow evolution at $Re = 700$ with $L = 1.955$

### 4 CONCLUSIONS

A new application of the global method of approximated particular solutions to solve the two-dimensional unsteady Navier-Stokes equations is presented. A first order implicit scheme is used for temporal discretization and the resultant system of non-linear equations is solved by the Newton-Raphson method with variable step. The accuracy of the numerical scheme is verified by using the Taylor’s decaying vortex problem. The com-
puted results are in good agreement with the analytical solution, showing a second-order convergence rate in terms of nodal spacing. The two-sided lid-driven cavity problem is solved and excellent results are obtained at $Re = 700$. The proposed scheme reproduces well the changes presented in the flow pattern in which a central recirculation zone is obtained if $1 < l < 2$ and three recirculation zones are observed when $2 \leq l$. These first satisfactory solutions of unsteady flow problems by MAPS are the base of future applications to the solution of more complex situation such as high-$Re$ and compressible flow.

REFERENCES


MODEL REDUCTION FOR SHALLOW WATER FLOWS

Matthew Farthing\(^1\), Chris Kees\(^1\) and Alexander Lozovskiy\(^2\)

\(^1\)Coastal and Hydraulics Laboratory
U.S. Army Engineer Research and Development Center
3909 Halls Ferry Road
Vicksburg, MS 39180-6199.
mwfarthing@gmail.com, cekees@gmail.com

\(^2\)Institute of Scientific Computation
Texas A&M University,
College Station, TX 77840.
lozovskiy@math.tamu.edu

Key words: Shallow Water equations, POD, DEIM, riverine applications

Abstract. To capture the hydrodynamics of shallow water flows in natural systems, it is often essential to characterize the system’s bathymetry as well as the impact of bottom stress caused by surface roughness. Unfortunately, constants like Manning’s friction coefficient in common parameterizations of bottom stress are not directly measurable, while direct measurement of bathymetry can be expensive and impractical in many environments. This leads to the need to solve one or more inverse problems to estimate depths and/or roughness parameters using available data, which may be limited to surface velocities and/or free-surface elevations. While a variety of inversion techniques can be used, most if not all can be expected to require a large number of (forward) shallow water simulations to provide an representative set of velocities and depths. These computations can be quite expensive and serve as a limiting factor in the effectiveness of the inversion process. In this work, we provide a global model reduction framework for stabilized finite element schemes for the shallow water equations (SWE) by means of Proper Orthogonal Decomposition (POD). An “optimal” reduced order basis is constructed based on high-dimensional solution snapshots. Additional nonlinear complexity reduction is made using the Discrete Empirical Interpolation method (DEIM). We present numerical results to compare accuracy and execution times to justify our choice for model reduction in the context of a forward simulation.

1 INTRODUCTION

With advances in numerical methods for transport equations, many techniques have been developed for the simulation of physical flows, including large-scale natural phenom-
ena such as rivers with complex bed topographies and bottom friction. In environmental flows, values for material properties and parameters given as input to the transport equation(s) often have strong impact on the resulting solution (output). Thus, it is crucial that this data be available with reasonable accuracy due to the direct effect on flow dynamics as well as secondary processes like transport of constituents and sediment, fish migration, and vessel trafficability [10]. Given this sensitivity, it is not surprising that there has been significant interest over the years in developing techniques for measuring or inferring properties like topography (or bathymetry).

In the context of an inverse problem, some information on a property like bathymetry or surface roughness is obtained based on flow characteristics. The motivation for inversion stems from the fact that direct measurements of properties like bottom elevation are typically more difficult (and expensive) than that of data at or near the free surface, while some properties like constants in common parameterizations of bottom stress are not directly measurable at all.

There are a host of inversion techniques from deterministic nonlinear least-squares approaches to ensemble-based techniques like ensemble Kalman filtering [6, 12] and more recent measure-theoretic methods [3]. While the theoretical underpinnings may differ widely, most if not all of these techniques can be expected to require many forward simulations to produce estimates of state variables like the velocity, $u$, or depth, $h$ for a range of possible values of parameters like the bottom elevation, $b$ or Manning’s $n$ [13, 14] that will vary as functions of spatial coordinates $x,y$ in the flow domain.

Such studies can require tens of thousands of simulations to fully explore the parameter space, particularly when the flow dynamics are highly nonlinear (as they are in shallow water systems) and the dimensionality of the inverse problem is high. Our possibilities, however, are limited in the sense that each realistic computation would require a relatively fine mesh. The resulting dynamical system, even on a parallel high performance system, can take hours to days for each simulation.

In order to be able to run many more than a few simulations, we propose global model reduction using POD. Using POD, high-fidelity models are replaced with reduced-order models to bridge this gap between physics-based modeling and engineering analysis. At a high level, our aim is to collect fine-scale snapshots of good accuracy using a small set of parameterizations of a material (or geometric) property like $b(x,y)$ as input during an offline (training) stage. With these snapshots, we then construct a basis for new modes using Singular Value Decomposition (SVD) or other similar techniques to represent coherent structures of the flow. After truncation, only the most significant modes are kept in the system, and the latter is now expected to be cheap to compute. These cheap computations are carried out during the online (analysis) stage.

The paper is organized as follows. In Section 2, we present general two-dimensional shallow water equations for depth-averaged, free-surface flow and then present a continuous Galerkin (CG) formulation for its solution. Unfortunately, a direct CG approximation creates non-physical oscillations in the solution, yielding it useless for practical applica-
m. farthing, c.kees and a. lozovskiy

...tion and comparison to collected data. to cure this, we present stabilization based on the variational multiscale method [7] and an algebraic subgrid scale (asgs) approximation, in which the solution at scales smaller than allotted mesh element width is approximated as a (linear) function of the total strong residual. to further control oscillations, the stabilized approximation is enhanced with isotropic shock-capturing. in section 3, a global model reduction technique via svd/pod is applied to the swe. there, the method will be briefly demonstrated in general, and afterwards, it will be applied directly to the discrete, stabilized swe scheme. additional approximation of the nonlinear terms in the scheme via deim are summarized in section 4. numerical tests, performed for a simple academic solution, are provided in section 5, where both accuracy and runtimes are examined.

2 fine scale formulation for shallow water equations

the general two-dimensional shallow water system in a conservative form, with coriolis force neglected, reads as

\[
\frac{\partial \mathbf{u}}{\partial t} + \nabla \cdot (\mathbf{F} - \mathbf{D} \nabla \mathbf{u}) + \mathbf{r} = 0, \tag{1}
\]

for unknowns \( u_1 = h, u_2 = hu, u_3 = hv \), where \( h \) is a total depth, \( u \) and \( v \) are velocity components in the \( x \)- and \( y \)-direction respectively. \( \mathbf{F}(\mathbf{u}) \) denotes 3 by 2 advection matrix

\[
\mathbf{F} = \begin{pmatrix} F_1 & F_2 \end{pmatrix} = \begin{pmatrix} hu & hv \\ \frac{(hu)^2}{h} + \frac{gh^2}{2} & \frac{(hv)h}{h} \\ \frac{(hu)h}{h} + \frac{gh^2}{2} & \end{pmatrix},
\]

\( \mathbf{D} = \nu \text{diag}(0, 1, 1) \) denotes diffusion matrix with turbulent viscosity \( \nu \), \( g \) denotes gravitational acceleration, \( \mathbf{r}(\mathbf{u}) \) denotes reaction containing both the topography gradient and the manning’s bottom friction with (bounded) coefficient \( n \)

\[
\mathbf{r} = gh \begin{pmatrix} 0 \\ \frac{\partial b}{\partial x} \\ \frac{\partial b}{\partial y} \end{pmatrix} + gn^2 \sqrt{(hu)^2 + (hv)^2} \begin{pmatrix} 0 \\ \frac{hu}{h} \\ \frac{hv}{h} \end{pmatrix}.
\]

the first scalar equation is a continuity equation and the other two form a vector momentum equation. for simplicity, we assume that the boundary conditions are of dirichlet type for discharge \( u_2 \) and \( u_3 \), with no specific information provided for depth \( u_1 \).

the classical weak form for (1) produces

\[
\int_{\Omega} \frac{\partial \mathbf{u}}{\partial t} \cdot \mathbf{w} - \int_{\Omega} \mathbf{F} \cdot \nabla \mathbf{w} + \int_{\Omega} \mathbf{D} \nabla \mathbf{u} \cdot \nabla \mathbf{w} + \int_{\Omega} \mathbf{r} \cdot \mathbf{w} + \int_{\partial \Omega} ((\mathbf{F} - \mathbf{D} \nabla \mathbf{u}) \cdot \mathbf{n}) \cdot \mathbf{w} = 0 \tag{2}
\]

with arbitrary test function \( \mathbf{w} \) from a suitable space.
The Dirichlet conditions for \( u_2 \) and \( u_3 \) impose zero trace on the test functions \( w_2 \) and \( w_3 \), so only the first component of the boundary term remains. We obtain
\[
\int_{\partial \Omega} \left( (F - D \nabla u) \cdot \mathbf{n} \right) \cdot \mathbf{w} = \int_{\partial \Omega} (hu \cdot n_x + hv \cdot n_y) w_1 = \int_{\partial \Omega} (u_2 \cdot n_x + u_3 \cdot n_y) w_1
\]
Therefore, the boundary term is treated as a mere input forcing term. System (2) produces a stable solution, i.e. with realistic non-oscillatory profile, as long as the mesh resolution is sufficiently small. For other cases, some form of stabilization is required.

If \( \phi_s \) denotes the \( s \)-th piecewise linear finite element shape function, then the discrete analog of solution \( u_i \) to be plugged into (2) is represented via coefficient vector \( v_i \) of size \( N \) as
\[
u_{h,i} = \sum_{s=1}^{N} \phi_s v_{i,s}, \quad 1 \leq i \leq 3. \tag{3}
\]
Here, we propose an ASGS method to stabilize the CG finite element discretization of the SWE. The foundational variational multiscale approach was thoroughly demonstrated by Hughes in [7] for abstract stationary problems with Dirichlet boundary conditions. We largely follow [8], which showed the method in a more practical way and applied it to porous media flow and one-dimensional shallow water system with flat topography and no bottom friction.

Let us introduce advection matrices \( A_i(u_h) \) and vector operator \( \mathcal{M}_{u_h} \) as
\[
A_i(u_h) = \frac{\partial F_i}{\partial u}(u_h) \tag{4}
\]
and
\[
\mathcal{M}_{u_h} w_h = -\sum_i A_i(u_h)^T \frac{\partial w_h}{\partial x_i} \tag{5}
\]
Here we denote \( x_1 = x, \ x_2 = y \). With \( a(\cdot, \cdot) \) denoting the form for the stationary part of (2), the stabilized grid scale equation takes the following form: \( \forall w_h \)
\[
\int_{\Omega} \frac{\partial u_h}{\partial t} \cdot w_h + a(u_h, w_h) + \sum_e \int_{\Omega_e} \mathcal{M}_{u_h} w_h \cdot \tau R(u_h) = 0, \tag{6}
\]
where \( \tau \) is a matrix of intrinsic time scales and \( R \) is a strong residual of (1).

The main issue is specification of \( \tau \). The derivation of the optimal \( \tau \) for the most general formulation (6) is a challenging task and is usually done in a somewhat ad hoc way. From here, we impose a separate \( \tau_i \) for each of the terms in the sum of (4), following the approach in [2].
\[
\int_{\Omega} \frac{\partial u_h}{\partial t} \cdot w_h + a(u_h, w_h) - \sum_i \int_{\Omega_e} R(u_h) \cdot \sum_j (A_i(u_h) \tau_j)^T \frac{\partial w_h}{\partial x_i} = 0 \tag{7}
\]
Each $\tau_i$ is obtained as optimal for a one-dimensional problem with the reaction term $r$ in $a(u, w)$ removed. In such a derivation, we use the basic methodology from the streamline upwind Petrov-Galerkin (SUPG) approach, obtaining an alternative to the value presented in [8] for SWE.

The ASGS stabilization (7) is expected to remove or control instability arising due to convection. It is, however, often not able to resolve the solution profile monotonically near possible shock fronts. The latter is usually treated via the addition of a shock-capturing term having the form of an artificial viscosity. Here, we use a simple isotropic approach as in [9]. The final form of our finite element scheme for fine scale simulations is

$$\int_{\Omega} \frac{\partial u_h}{\partial t} \cdot w_h + a(u_h, w_h) - \sum_e \int_{\Omega_e} R(u_h) \cdot \sum_i (A_i(u_h) \tau_i)^T \frac{\partial w_h}{\partial x_i} +$$

$$\sum_e \int_{\Omega_e} c_e \frac{h_e}{2} \left| \nabla u_h \right| \text{diag}(0, 1, 1) \nabla u_h \cdot \nabla w_h = 0 \quad (8)$$

**2.1 Temporal discretization**

For marching in time with time step $\Delta t$, a Backward Euler approximation is proposed (higher order backward difference schemes can be used to increase temporal accuracy). In order to avoid solving a heavy nonlinear algebraic system at each time step, a semi-implicit modification of the above scheme is proposed in such a way that extrapolation is introduced inside all nonlinear terms in (8). This lagging does not influence the accuracy of the scheme, but makes it only conditionally stable. The extrapolation in the ASGS stabilizing term is performed as

$$-\sum_e \int_{\Omega_e} R(u_h^k) \cdot \sum_i (A_i(u_h^{k-1}) \tau_i^{k-1})^T \frac{\partial w_h}{\partial x_i}, \quad (9)$$

when solving for time level $k$. Additional extrapolation in $R(u_h^k)$ itself is used.

In the shock-capturing term, we remove all the implicitness from the artificial viscosity coefficient. Due to inability to approximate time-derivative term in $R(u_h)$ with only one previous time level $k - 1$, the extrapolation for the shock-capturing term involves two previous time levels $k - 1, k - 2$. This lagging time discretization of the shock-capturing term looks like

$$\sum_e \int_{\Omega_e} c_e \left( \frac{u_h^{k-1} - u_h^{k-2}}{\Delta t} + S(u_h^{k-1}) \right) \text{diag}(0, 1, 1) \nabla u_h^k \cdot \nabla w_h$$

with $S(u_h)$ denoting the stationary part of the residual.

At each time step, the resulting linear system to be solved has form

$$\begin{pmatrix}
A_{11} & A_{12} & A_{13} \\
A_{21} & A_{22} & A_{23} \\
A_{31} & A_{32} & A_{33}
\end{pmatrix}
\begin{pmatrix}
v_1 \\
v_2 \\
v_3
\end{pmatrix} =
\begin{pmatrix}
f_1 \\
f_2 \\
f_3
\end{pmatrix} \quad (10)$$
Each block $A_{ij}$ corresponds to test function component $w_i$ and solution component $u_j$, and $f_i$ is the inertial contribution of the solution from the previous time level including the ASGS stabilization terms. Each block is viewed as a sum of linear $L$, i.e. solution independent, and nonlinear block terms $N$, i.e. those depending on the solution at previous time levels. If we neglect Manning’s friction, then component $k_l$ of each block $A_{ij}$ is evaluated as
\[
\{A_{ij}\}_{kl} = \{L_{ij}\}_{kl} - \int_{\Omega} \phi_l c_{ij} \cdot \nabla \phi_k d\Omega \\
\int_{\Omega} \nabla \phi_l \cdot B_{ij} \nabla \phi_k d\Omega + \frac{1}{\Delta t} \int_{\Omega} \phi_l m_{ij} \cdot \nabla \phi_k d\Omega + \int_{\Omega} d_{ij} \nabla \phi_l \cdot \nabla \phi_k d\Omega,
\]
(11)
The first term on the right hand side (RHS) is based on the linear part (2). The second term in the RHS corresponds to the flux term in (2) and is the only physics-based nonlinearity. All the following nonlinear terms emanate from numerical stabilization.

3 PROPER ORTHOGONAL DECOMPOSITION

The fine scale solver presented above allows us to perform simulations producing a stabilized profile of the solution, whose accuracy is consistently improved with mesh refinement. It is assumed the mesh for the offline stage is taken as such that resolves as many features of the solution as the computational resources allow within maximally affordable time. Such precision is only required for accumulating the snapshot set in order to construct a sufficiently wide basis space. For online simulations — intended to run as many times as needed to perform the inversion — a much cheaper reduced order model is required.

POD is an approximation method of global type that has been applied to general large-scale parameterized dynamical systems. It is based on the representation of the high-dimensional state obeying the governing evolution equations by an element of a low-dimensional subspace that is chosen based on energetic considerations. The second part of the method involves projecting the equations themselves onto that chosen subspace, yielding a dynamical system of reduced order. We recommend the work in [4] for a good demonstration of the POD methodology.

Recalling (3), let $S_i = (v_{i1}^1, ..., v_{iM}^M)$ be snapshots obtained from offline runs from $t = 0$ to $t = T$ for each of the three variables $u_i$. The set $S_i$ typically consists of snapshots at intermediate values of $0 < t < T$, for each corresponding input parameter or forcing function. The snapshots themselves are not suited as a basis, since the set may be poorly conditioned or even of incomplete rank. Instead, an SVD is proposed:

$$S_i = U_i \Sigma_i V_i^T, \quad 1 \leq i \leq 3.$$ 

with $\Sigma_i = \text{diag}(\sigma_i^1, ..., \sigma_i^M)$ being the diagonal matrix containing the positive eigenvalues in decreasing order. It is expected that the eigenvalues will decrease to zero rapidly from the
first few being several orders larger than the rest. This is desirable for model reduction, since “energetically” the majority of the information in the original solution can then be reconstructed within only a low-dimensional space. The columns of the matrix $U_i$ provide the desired basis vectors, ordered correspondingly with $\sigma_i$.

Denoting the reduced space basis as

$$\tilde{U}_i = U_i \begin{pmatrix} I_{m_i \times m_i} & 0 \end{pmatrix}$$

the reduced order discrete solution $z_i$ of size $m_i \ll N$ is connected to the high-fidelity discrete solution via approximation

$$v_i \approx \tilde{U}_i z_i \quad (12)$$

Plugging (12) into (10) and multiplying from the left the corresponding block row by $U_i^T$ (Galerkin projection), gives

$$\begin{pmatrix} \tilde{U}_1^T A_{11} \tilde{U}_1 & \tilde{U}_1^T A_{12} \tilde{U}_2 & \tilde{U}_1^T A_{13} \tilde{U}_3 \\ \tilde{U}_2^T A_{21} \tilde{U}_1 & \tilde{U}_2^T A_{22} \tilde{U}_2 & \tilde{U}_2^T A_{23} \tilde{U}_3 \\ \tilde{U}_3^T A_{31} \tilde{U}_1 & \tilde{U}_3^T A_{32} \tilde{U}_2 & \tilde{U}_3^T A_{33} \tilde{U}_3 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} \tilde{U}_1^T f_1 \\ \tilde{U}_2^T f_2 \\ \tilde{U}_3^T f_3 \end{pmatrix} \quad (13)$$

This system of order $m_1 + m_2 + m_3$ must be solved on each time level to provide low-dimensional solution $z_1$, $z_2$, $z_3$. The assembly of the matrix from (13) is performed at each time step. We first reconstruct the fine-scale solution using (12) and then reassemble the high-order matrices $A_{ij}$ using extrapolation as discussed above. These two operations are costly since they are conducted on the original high-dimensional space.

4 DISCRETE EMPIRICAL INTERPOLATION METHOD

The blocks $L_{ij}$ of $A$ related to linear terms in (11) are independent of time and can be reduced once in (13) before temporal evolution of the system. Other terms, however, require recomputation in the original high-dimensional space and subsequent reduction back to the reduced space at every time step. Such computational burden, involving operations with complexity that polynomially depends on the order of the high-dimensional space, annuls the execution speed achieved by POD reduction.

To overcome this performance issue, further complexity reduction — this time applied to the nonlinearity — is proposed via the Discrete Empirical Interpolation Method (DEIM) [4]. DEIM is the discrete analog of the Empirical Interpolation method [1] and has been widely used, for example, in the context of reservoir simulations [5] and has also been applied to shallow water flows in the context of an implicit ADI finite difference scheme [11].

If $f$ denotes a vector with large dim$f$ and $V$ denotes basis with many fewer orthonormal vectors to which the projection of $f$ is expected to be done, then we expect, assuming reasonable accuracy,

$$f \approx Vc$$
for some coefficient vector $c$. The best possible approximation for a fixed vector $f$ in the standard 2-norm is a projection $VV^T f$. This is, however, an expensive procedure. Instead, we propose interpolation.

Interpolation with a given finite set of indices by its nature evaluates the vector at those indices exactly, but is approximate at other indices. Let $P$ denote an index selection matrix, i.e. it is filled with zeros and ones in such a way, that each column has one only at the position of interpolation index. Then the following is exact:

$$P^T f = P^T V c.$$

From here, we approximate our function $f$ through $f_{\text{approx}}$ via

$$f_{\text{approx}} \approx V (P^T V)^{-1} P^T f = QP^T f.$$

If the number of interpolation indices is way below $\text{dim} f$, the evaluation of $P^T f$ should be treated not as a brute-force matrix-by-matrix multiplication, but as a selection of only specific positions from $f$ for evaluation, granting complexity that directly depends only on the quantity of interpolation indices, which is the same as the quantity of vectors in basis $V$.

For the SWE system (13), we use $V_i$ generated through the usual SVD, applied to the snapshots of the nonlinear terms collected during the offline stage, where $i$ corresponds to the number of equation in (1). Due to extrapolation in time at time level $k$, each $i$-th nonlinear term, reduced thanks to POD, is viewed as $\hat{U}_i^T \sum_j N_{ij} (\hat{U} z_{k-1}, \hat{U} z_{k-2}) \hat{U}_j z_{j,k}$. The interpolation in this case is performed as

$$\tilde{U}_i^T Q_i P_i^T \sum_j N_{ij} (\hat{U} z_{k-1}, \hat{U} z_{k-2}) \hat{U}_j z_{j,k}.$$

In this case, term $\tilde{U}_i^T Q_i$ is precomputed only once. The interpolation via matrix $P_i$ has to be viewed as mere selection of a small set of test shape functions $w_h$ in the variational form of (8) when evaluating nonlinear terms of the dynamical system at each time step. In addition to the speed-up provided by efficient evaluation of $P_i^T N_{ij}$, better performance can be achieved thanks to sparsity structure of matrices $N_{ij}$. For each fixed row of $N_{ij}$, only neighboring nodes contribute to its computation. Therefore, with information on mesh connectivity available, complexity of the operation $\sum_j P_i^T N_{ij} \hat{U}_j$ can be significantly reduced. We can also narrow our computational efforts only to those rows in $u_h = \hat{U} z$ that are required for evaluation of nonlinear term $N_{ij} (\hat{U} z_{k-1}, \hat{U} z_{k-2})$, since others play no role except when used during data collection for non-computational post-processing purposes.

DEIM itself is responsible for choosing the “optimal” set of indices in $P$. The index selection is done during the pre-computational phase and only requires set of vectors $V$ as input. More on DEIM indices selection algorithm can be found in [4].
5 Numerical tests

Initial numerical tests were conducted on the rectangular domain $[0, 10] \times [0, 1]$ with flat topography $\nabla b = 0$ and no friction, the initial condition shown in Figure 1 was $u_1 = 1 + e^{-4(x-2)^2}$ and $u_2 = u_3 = 0$. Boundary conditions were $u_2 = u_3 = 0$ on $\partial \Omega$, viscosity $\nu = 10^{-4}$ and time step $\Delta t = 10^{-3}$. A purely uniform triangular mesh containing $N = 6526$ nodes was used. Figure 2 presents the solution at $T = 1$.

![Figure 1: Initial data](image1)
![Figure 2: Fine grid solution at $T = 1$](image2)

5.1 POD

Results for a full POD approximation, with the same input parameters as above, are shown in Figure 3, while the error is given in Figure 4. With 50 modes (out of 1000 collected during fine-grid offline stage) for both $u_2$ and $u_3$, the POD solution was tested for different number of modes for depth, $u_1$, from 3 to 50. The wave front was recovered well by POD even with only few modes present in the system, although the profile showed smaller scale “wavy” oscillations for up to 40 modes due to the increasing frequency of the basis functions.

The relative error $\|u_i - u_i^{POD}\|_2/\|u_i\|_2$, $1 \leq i \leq 3$, is presented in Figure 4. We clearly see decay of the error as we increase the number of modes even just for $u_1$. After sufficiently many modes, the error tends to remain the same, as more modes for $u_2$ and $u_3$ have to be introduced into the system in order to further reduce the error. Note that even though the relative error for $u_3$ tends to be higher than that for other unknowns, the norm $\|u_3\|_2$ itself is negligible.

The relative runtime $T^{POD}/T$, where $T$ is the runtime for fine grid solver, increased linearly with respect to the number of modes for $u_1$, from 1.0071 for 3 modes, to 1.1088 for 50 modes. This clearly indicates lack of advantage when using a pure POD approach as opposed to the original fine-grid solver. The reduction process of the matrix in (13) mainly contributed to the loss of the speed-up obtained by the solution process of the reduced system.
Figure 3: Solution profile $u_1$ at $T = 1$, with 50 modes both for $u_2$ and $u_3$

Figure 4: Relative $L^2$-error between fine grid and POD solutions for all $u_i$ at $T = 1$, with 50 modes both for $u_2$ and $u_3$
5.2 POD-DEIM

The results of the POD-DEIM solver are summarized in Figure 5. The POD-DEIM solution was compared to the full POD approximation from the previous subsection with 50 modes used for all three $u_i$. Fixing 100 interpolation points for $u_2$ and $u_3$, the number of interpolation points for depth $u_1$ was increased from 2 to 100. The resulting solution appears almost insensitive to the order of interpolation, showing only negligible drop of relative error within the first 50 or so interpolation points. Such behavior is beneficial for forward simulations, since the same accuracy is obtained with much less computational effort. The runtime for DEIM also increased linearly with respect to the order of interpolation. The range of ratio $T_{DEIM}^\text{POD}$ was from 0.0337 for 2 points to 0.0547 for 100 points, which is a significant acceleration.

![Figure 5: Relative $L^2$-error between POD and POD-DEIM solutions for all $u_i$ at $T = 1$, with 100 DEIM interpolation points for $u_2$ and $u_3$ and 50 modes for all $u_i$](image)

REFERENCES


VENTILATION SYSTEM FOR DRYING OUT WALLS AFTER A FLOOD: INFLUENCE OF THE THICKNESS OF THE WALLS

ISABEL TORRES¹ AND ÓSCAR LÓPEZ²

¹Department of Civil Engineering – University of Coimbra
Rua Luis Reis Santos – Pólo II – 3030-788 Coimbra - PORTUGAL
Email: itorres@dec.uc.pt

²Institute for Research and Technological Development in Construction Sciences
Rua Pedro Hispano, s/n – Pólo II – 3030-289 Coimbra - PORTUGAL
Email: oscar.lopez@itecons.uc.pt

Key words: Historical buildings, flood, ventilation system, numerical simulation.

Abstract. The conservation of historical buildings assumes, nowadays, a considerable importance and it has had a great development in these past few years. We can say that among all the different kinds of pathologies, dampness in historical buildings is an important one. The most emblematic buildings, for historical reasons, were built along the floodplains presenting, therefore, an increased exposure to the probability of a flood. After the occurrence of a flood building elements will see its moisture content suddenly increased, with all losses resulting therefrom. It is very important that, after the occurrence of a flood, the reduction in moisture level of the walls is achieved as quickly as possible. In this past years, research was carried out in order to develop a method to minimize the effects of rising damp. This method consists of installing ventilated systems at the wall base. This system has been validated experimentally and numerically and has achieved satisfactory results in what concerns to the reduction of the level achieved by the water front in the wall. Now we are performing numerically investigation to validate the efficiency of a base wall ventilation system as a technique to improve the drying of walls of older buildings after a flood.

Since experimental investigations are rather expensive and time-consuming, there is an increasing interest in use calculative methods in order to assess moisture behavior of building components. Computational programs allow a two-dimensional analysis of the simultaneous heat and moisture transport in building components, which is the best approximation to drying of moisture.

Our main goal is modelling the drying process of a wall after being affected by a flood, analyzing the efficiency of the ventilation system and analyzing the influence of different parameters on all the process.

What is intended to present in this paper are some of the results obtained that are part of a wide investigation that is being carried out. In this study we focused on the influence of wall thickness to determinate the efficiency of the ventilation system.
1 INTRODUCTION

Moisture content is one of the major causes of pathologies of Portuguese historical buildings. Mainly due to easier water supply reasons, our most prestigious buildings were built near natural water ways or along floodplains and thereby they are exposed to a high probability of flooding [1].

During a flood building elements come in direct contact with water increasing their moisture content quickly. The most threatened building elements are foundations and wall bases, which generally have difficulty to dry for a natural process because the weather conditions are not favourable and/or because they are elements with high thicknesses which hinders the removal of moisture content. This phenomenon is more damaging as more events occurs during the year.

Floods are occasional events however, in many cases, buildings may have other sources of moisture that may increase its moisture content. One of the most important and prejudicial is rising damp because it can be a continuous source of moisture throughout all the year.

The phenomenon of rising damp describes the movement of moisture upward through permeable building materials by capillary action. One of the most common sources of rising damp is due to high groundwater levels and/or a poor surface drainage in the base of the walls of buildings. It becomes a problem if the moisture penetrates vulnerable materials or finishes, particularly in the occupied parts of a building. Additionally, this moisture will dissolve soluble salts from the building materials such as calcium sulphate and may also carry soluble salts from the ground.

Over the last years, research has been carried out, at the Department of Civil Engineering of the University of Coimbra in collaboration with the Department of Civil Engineering of the University of Porto to develop a wall base ventilation system to minimise the effects of the phenomenon of rising damp [2]. In the Institute for Research and Technological Development in Construction Sciences (ITeCons) in collaboration with the Department of Civil Engineering of the Faculty of Engineering of the University of Coimbra a large numerical investigation is being development to use this wall base ventilation system as a technique to improve and accelerate the drying process of walls in historical buildings after a flood (Figure 1) [3].

Figure 1: Ventilation system for drying out walls after a flood
2 VENTILATION SYSTEM FOR DRYING OUT BUILDINGS

2.1 The technique

The ventilation system for drying out walls after a flood consists of circulating air at the base of the walls through peripheral channels in contact with them. With these channels we prevent lateral infiltrations and increase the evaporation and thereby reduce the moisture content of the wall. To increase the circulation of the air into these channels hydro-regulated mechanical ventilation devices (extractors) can be introduced [3].

![Figure 2: Principle of ventilation system for drying out buildings after a flood: (a) moisture content increase due to infiltrations in the sequence of a flood, (b) drying out by wall base ventilation system and by natural drying process](image)

3 NUMERICAL SIMULATION

3.1 Calculation program used

Nowadays, calculation programs are essential instruments for analyzing the hygrothermal behavior of building components under certain climatic environments. In order to simulate the drying process of a wall after a flood we used WUFI-2D calculation program developed in Fraunhofer Institute for Buildings Physics (3.3 version). This software has already been used to perform numerical simulations to validate the wall base ventilation system to treat rising damp mentioned above and developed at the University of Coimbra and University of Porto [2, 4, 5].

WUFI-2D is based on the following heat and moisture transfer equations [6, 7, 8]:

\[
\frac{dH}{dt} \frac{\partial T}{\partial t} = \nabla (\lambda \nabla T) + h_v \nabla (\delta_p \nabla (\Phi p_{sat}))
\]  

(1)

\[
\frac{dw}{d\Phi} \frac{\partial \Phi}{\partial t} = \nabla (D_{\phi} \nabla \Phi) + \delta_p \nabla (\Phi p_{sat})
\]  

(2)

Where \(\frac{dH}{dT}\ [J/m^3]\) is the heat storage capacity of the moist building material, \(\frac{dw}{d\Phi}\ [kg/m^3]\) the moisture storage capacity of the building material, \(\lambda\ [W/m K]\) the thermal...
conductivity of the moist building material, $D_\phi \text{[kg/m s]}$ the liquid conduction coefficient of the building material, $\delta_p \text{[kg/m s Pa]}$ the water vapour permeability of the building material, $h_v \text{[J/kg]}$ the evaporation enthalpy of the water, $p_{sat} \text{[Pa]}$ the water vapour saturation pressure, $T \text{[°C]}$ the temperature and $\Phi$ [-] is the relative humidity [6, 7, 8].

In many cases simulation programs require the introduction of many simplifications in order to the behavior of the numerical model be similar to the reality. In this regard, the 3.3 version of WUFI-2D allows the use of “air change sources”. The “air change source” resource permits us to simulate the renovation of the air into channels. Depending on the temperature and humidity conditions in the component and in the outdoor or indoor air, heat and moisture may be transported into or out of the component. We have used this resource because we believe it is the best way to simulate the characteristics of the inside air of the wall base ventilation system because it has not infinite evaporation capacity and it depends on the characteristics of the air (temperature and relative humidity).

3.2 Analyzed parameters

In this study we analyzed the influence of the wall thickness on the efficiency of the base wall ventilation system for drying out walls after a flood and in the presence of rising damp. We considered a wall of 2.00 meters high. The ventilation system for drying out consists in two channels with 0.20x0.20 meters placed on both sides of the wall base. A waterproofing between sand and the ventilation system was considered to prevent lateral infiltrations into the channels. Maintaining the height and the dimensions of the channel constants we changed the thickness of the wall from 0.20 m to 1.00 m. Figure 3 shows the five geometry configurations studied.

<table>
<thead>
<tr>
<th>Configuration 1</th>
<th>Configuration 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without ventilation system</td>
<td>With ventilation system</td>
</tr>
<tr>
<td>0.20 meters thick</td>
<td>0.40 meters thick</td>
</tr>
</tbody>
</table>

**Figure 3**: Studied configurations to analyze the influence of the wall thickness (continued on next page)
The use of numerical simulation programs requires a deep knowledge of the characteristics of the materials and of boundary conditions. In our wall we have used the materials’ hygrothermal properties that were experimentally obtaining during the validation of the ventilation system applied to calcareous walls to treat the rising damp (Table 1) [4].

Table 1: Hygrothermal properties of the material

<table>
<thead>
<tr>
<th>Hygrothermal property</th>
<th>Calcareous stone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulk density [kg/m³]</td>
<td>2155.0</td>
</tr>
<tr>
<td>Heat capacity [J/kg K]</td>
<td>1000</td>
</tr>
<tr>
<td>Open porosity [%]</td>
<td>18.8</td>
</tr>
<tr>
<td>Thermal conductivity [W/m K]</td>
<td>1.33</td>
</tr>
<tr>
<td>Vapour diffusion resistance factor</td>
<td>29.4</td>
</tr>
<tr>
<td>Water absorption coefficient [kg/m² √s]</td>
<td>0.024</td>
</tr>
<tr>
<td>Free water saturation [kg/m³]</td>
<td>188</td>
</tr>
</tbody>
</table>
In order to simulate the drying process of a wall we began the simulations considering that the wall was totally wetted, that is to say that it had a relative humidity of 100%. In all simulations we considered the same climatic conditions: on one side of the wall we applied a real outdoor climate (from Lisbon) with data of wind, rain, barometric pressure, temperature, relative humidity and solar radiation (it is considered, based on the wind speed and its direction, the possibility that the wall absorbs rainwater) and on the other side we considered an indoor climate.

At this first stage we have considered that the simulation time was only one year. Provably, higher wall thickness will require longer drying times because we considered that the flood saturated the material. However, depending on the duration of the flood and the porosity of the building materials, the wall may not reach 100% of relative humidity throughout its thickness and the drying process may be faster because the most difficult moisture to remove is the water content at the vertical central profile.

In a previous study we analyzed the influence of the air change rate ($ACH$) of the ventilation cavity (as an air change source in WUFI-2D) in the drying process because it is one of the most important and decisive parameter that influences the potential evaporation of external air circulating into channels ($ACH$, in h$^{-1}$, is the air changes into a channel per hour) [3]. This study was performed for a thickness of 0.20 m and it corresponds to the first geometric configuration on figure 3. It was concluded that, for this configuration, there is an optimum $ACH$ into channels with which the maximum efficiency to reach the drying out of the wall is achieved. This study was performed for an air change rates up to 100 h$^{-1}$, that, in accordance to [5] and [3], are expectable rates for a length of wall base ventilation systems between 50 and 100 meters considering a maximum admissible air velocity of 3 m/s. In this way, $ACH$ can be calculated from equation (3) [3]:

$$ACH = \frac{Q}{V} = \frac{A \cdot u}{A \cdot L} = \frac{u}{L}$$

Where $Q$ [m$^3$/h] is the air flow rate, $V$ [m$^3$] is the volume of the channel, $A$ [m$^2$] is the cross-sectional area of the channels, $u$ [m/h] is the air velocity into channel and $L$ [m] is the length of the channel [3].

Maintaining all the other conditions apart from the thickness, our main goal was to analyze the influence of the of the wall thickness varying the air change rate (we considered values of $ACH$ of 1, 20, 40, 60, 80 and 100 h$^{-1}$). Thus, thirty-five numerical simulations were performed for the configurations presented in figure 3.

### 3.3 Results and discussions

WUFI-2D provides the water content, relative humidity, temperature and vapor pressure over time for all elements of the grid.

On figure 4 we present the evolution of the water content in the wall during one year of simulation for the five configurations studied and varying, in each one, the air change rate from 1 to 100 h$^{-1}$.
Isabel Torres and Óscar López

Figure 4: Variation of water content in the walls with ACH from 1 to 100 h\(^{-1}\) for the five configurations during one year.

Table 2 presents the variation of the water content for the different wall thickness and different air change rate after one year of simulation. Table 3 presents the water content reduction after one year of simulation and table 4 presents the relative efficiency of the ventilation system considering that the efficiency of 100% is obtained for the wall of 0.2 thick and a value of ACH of 100 h\(^{-1}\).

**Table 2**: Total water content after one year of simulation for all the walls and all air change rates of the ventilation channels

<table>
<thead>
<tr>
<th>Thickness of the wall (m)</th>
<th>Without ventilation</th>
<th>Water content (kg/m(^3))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 h(^{-1})</td>
<td>20 h(^{-1})</td>
</tr>
<tr>
<td>0.20</td>
<td>68.7</td>
<td>67.8</td>
</tr>
<tr>
<td>0.40</td>
<td>89.6</td>
<td>88.9</td>
</tr>
<tr>
<td>0.60</td>
<td>105.0</td>
<td>104.5</td>
</tr>
<tr>
<td>0.80</td>
<td>116.4</td>
<td>116.0</td>
</tr>
<tr>
<td>1.00</td>
<td>125.2</td>
<td>124.9</td>
</tr>
</tbody>
</table>
### Table 3: Water content reduction after one year of simulation for all the walls and all air change rates of the ventilation channels

<table>
<thead>
<tr>
<th>Thickness of the wall (m)</th>
<th>Water content reduction by introduction of the ventilation system (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 h⁻¹</td>
</tr>
<tr>
<td>0.20</td>
<td>1.3</td>
</tr>
<tr>
<td>0.40</td>
<td>0.8</td>
</tr>
<tr>
<td>0.60</td>
<td>0.5</td>
</tr>
<tr>
<td>0.80</td>
<td>0.4</td>
</tr>
<tr>
<td>1.00</td>
<td>0.2</td>
</tr>
</tbody>
</table>

### Table 4: Efficiency of ventilation system for all the walls and all air change rates compared with simulation with 0.20 meters thick and $ACH$ of 100 h⁻¹

<table>
<thead>
<tr>
<th>Thickness of the wall (m)</th>
<th>Efficiency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 h⁻¹</td>
</tr>
<tr>
<td>0.20</td>
<td>5</td>
</tr>
<tr>
<td>0.40</td>
<td>3</td>
</tr>
<tr>
<td>0.60</td>
<td>2</td>
</tr>
<tr>
<td>0.80</td>
<td>1</td>
</tr>
<tr>
<td>1.00</td>
<td>1</td>
</tr>
</tbody>
</table>

On tables 5 and 6 we present the water content profiles for the five geometry configurations without ventilation system, and with ventilation system and air change rate of 60 h⁻¹, respectively.

### Table 5: Water content variation along the transversal section of the walls of calcareous stone without ventilation system after one year of simulation

<table>
<thead>
<tr>
<th>Thickness of the wall</th>
<th>Water content (kg/m³)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.20 m</td>
<td></td>
</tr>
<tr>
<td>0.40 m</td>
<td></td>
</tr>
<tr>
<td>0.60 m</td>
<td></td>
</tr>
<tr>
<td>0.80 m</td>
<td></td>
</tr>
<tr>
<td>1.00 m</td>
<td></td>
</tr>
</tbody>
</table>

### Table 6: Water content variation along the transversal section of the walls of calcareous stone with ventilation system and air change rate of 60 h⁻¹
ventilation system and $ACH$ of 60 h$^{-1}$ after one year of simulation

<table>
<thead>
<tr>
<th>Thickness of the wall</th>
<th>0.20 m</th>
<th>0.40 m</th>
<th>0.60 m</th>
<th>0.80 m</th>
<th>1.00 m</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water content (kg/m$^3$)</td>
<td>0&lt;=x&lt;33</td>
<td>33&lt;=x&lt;67</td>
<td>67&lt;=x&lt;100</td>
<td>100&lt;=x&lt;133</td>
<td>133&lt;=x&lt;167</td>
</tr>
</tbody>
</table>

On figure 5 we present the water content at the central vertical section of the wall after one year of drying, for the five configurations and for the six different air change rates.

On Table 7 we can see the maximum percentage of water removed from the wall at a vertical central profile compared with the water content of the simulation without ventilation system.

**Figure 5:** Variation of water content in the wall with $ACH$ from 1 to 100 h$^{-1}$ for the five configurations during one year.
Table 7: Maximum percentage of water content removed from the wall at a vertical central section compared with the water content of the simulation without ventilation system

<table>
<thead>
<tr>
<th>Thickness of the wall (m)</th>
<th>Maximum percentage of water content removed from the wall at a vertical central profile (%)</th>
<th>Air change rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1 h⁻¹</td>
</tr>
<tr>
<td>0.20</td>
<td></td>
<td>5.6</td>
</tr>
<tr>
<td>0.40</td>
<td></td>
<td>2.3</td>
</tr>
<tr>
<td>0.60</td>
<td></td>
<td>1.1</td>
</tr>
<tr>
<td>0.80</td>
<td></td>
<td>0.6</td>
</tr>
<tr>
<td>1.00</td>
<td></td>
<td>0.4</td>
</tr>
</tbody>
</table>

As expected, as the wall thickness increases the total water content after one year of simulation increases. In fact, thicker walls need more time and favorable climate conditions to reach the drying out.

Although the moisture removed by the ventilation system, in kg/m³, decreases with increasing thickness (Figure 6), the moisture content removed from the wall, in kg, in the five configurations are the same (Figure 7). This means that the ventilation system reaches its maximum efficacy and the renovation air cannot evaporate more moisture content from the wall.

Figure 6: Water content removed from the wall by the ventilation system, in kg/m³, depending on the value of ACH, after one year of simulation

Figure 7: Water content removed from the wall by the ventilation system, in kg, depending on the value of ACH, after one year of simulation
When the wall thickness is small the influence of the ventilation system is higher. It can even be observed an optimum $ACH$ value from which the slope of the curve decreases. We see that from this point increasing the $ACH$ value does not involve a significant increase on the water content removed by the ventilation system. When the wall thickness increases this optimum $ACH$ value cannot be observed in the curves and the higher is the $ACH$ value the higher is the moisture content removed from the wall. This is due to the fact that for higher thicknesses the phenomenon of rising damp will introduce more moisture into the base of the wall them the flood because the level achieved by the moisture front is higher (the same happens with materials with high absorption coefficients).

On tables 5 and 6 we can see that, as expected, by increasing the thickness of the wall we are increasing the absorption by capillarity and in order to restore the equilibrium between the “absorbed flow” and the “evaporated flow” the moisture front must increase. Other previous studies focused on the influence of the thickness of the walls and their properties on the treatment of rising damp in historic buildings, without flood, also obtained the same conclusions [9].

The only way to increase further the evaporation is increasing the value of $ACH$ through decreasing the length of the ventilation system. Another option is to increase the sizes of the channels to reduce the level achieved by the moisture front due to rising damp [10] but often this is not possible due to limitations of execution of these channels.

12 CONCLUSIONS

The main conclusions of our study, which set out to analyze the influence of the thickness of the wall on a hygro-regulated ventilation system at the base to dry out buildings after a flood are:

- Ventilation of the base of the walls is a simple technique for improving the drying out of buildings after a flood that offers great potential to protect architectural heritage.
- Wall base ventilation system is a simple technology for decreasing the level achieved by the moisture front of rising damp.
- Varying the ventilation air change rate has a great influence on the efficiency of the wall base ventilation when the walls have a slow thickness, but has less influence on high thickness walls. We can see that for walls of 0.20 meters thick the water content reduction with the introduction of the ventilation system with an $ACH$ value of 100 h$^{-1}$ is 25.0%, for the wall of 0.40 meters the water content reduction is 10.1%, for the wall of 0.60 meters the water content reduction is 5.6%, for the wall of 0.80 meters the water content reduction is 3.8% and for the wall of 1.00 meters the water content reduction is 2.7%.
- The drying efficacy of the ventilation system at a vertical central profile of the wall decreases with the increase of wall thickness. We can see that for the wall of 0.20 meters thick the maximum decrease of the water content when we introduce the ventilation system is 60.0%, for the wall of 0.40 meters the decreased is 25.1%, for the wall of 0.60 meters the decreased is 11.7%, for the wall of 0.80 meters is 6.4% and for the wall of 1.00 meters is 3.9% after one year of simulation and for $ACH$ value of 100 h$^{-1}$. 


The main characteristics which determine the moisture of materials and thereby the drying process of a wall after a flood in presence of rising damp are the open porosity, the absorption coefficient and the absorption curve. Therefore, the next step will be to study the efficiency of the ventilation system using materials with different characteristics, including one of the most common materials in the Portuguese historical buildings, Ançã’s stone, which is being characterized specifically for this project.

ACKNOWLEDGMENTS

The authors would like to thank the FCT – Fundação para a Ciência e Tecnologia for its support through the project PTDC/ECM-COM/3080/2012 Development and optimization of a hygro-regulated system for drying out buildings after a flood. This work has been framed under the Initiative Energy for Sustainability of the University of Coimbra and supported by the Energy and Mobility for Sustainable Regions - EMSURE - Project (CENTRO-07-0224-FEDER-002004).

REFERENCES

MODELLING THE AIR-FUEL MIXTURE IN AN ATMOSPHERIC BURNER

Ana Carolina Sá², Senhorinha Teixeira¹ and José C. Teixeira²

¹Department of Production and Systems, School of Engineering, University of Minho
Azurém, 4800-058 Guimarães, Portugal
st@dps.uminho.pt

²Department of Mechanical Engineering, School of Engineering, University of Minho
Azurém, 4800-058 Guimarães, Portugal
jt@dem.uminho.pt

Key words: Atmospheric element burner, CFD simulation, excess air factor.

Summary. Atmospheric burners are commonly applied for domestic water heating purposes in European paradigm. It is well known hydrocarbon fuels combustion leads to GreenHouse Gases (GHG) production, such as nitrogen oxides (NOₓ). As the EU policies require at least a 40% domestic reduction in GHG emissions by 2030, compared to 1990, gas instantaneous water heater appliances must address a performance – pollutant emissions commitment. Some of the high-end units control the pollutant emissions by pre-mixing the whole combustion air with the fuel and in the process yield low emissions. However the vast majority of the burners operate on a simple natural convection draft of the combustion air that does not guarantee an accurate air to fuel ratio. The optimization of such appliances may prove a good compromise between the cost and environmental efficiency. The present paper reports the development of a computational model to describe the air/fuel mixture inside a dual stage atmospheric burner that operates on the principle of a rich/lean mixture. The results show that the air fuel ratio of the two stages promote a good fuel mixing with different richness levels.

1 INTRODUCTION

It is well known that hydrocarbon fuels combustion leads to GreenHouse Gases (GHG) production, such as Nitrogen Oxides (NOₓ). It is claimed that the harmful effects of the latter pollutant emissions into the atmosphere is the reason why so many researchers have been addressing the formation and control processes issues on what NOₓ is concerned [1-7]. To our knowledge, despite industrial burners have been frequently been taken into account, eg. [6, 8-11], less effort has been conducted in the domain of domestic water heater burners. Atmospheric burners are commonly applied for domestic water heating purposes is an European paradigm. According to Krafta et al. [12], domestic appliances are responsible for 6% of total NOₓ emissions. Regarding EU policies [13] that mandates at least 40% domestic reduction in GHG emissions by 2030, compared to 1990, gas instantaneous water heater appliances must address a performance – pollutant emissions commitment.

Excess air and temperature of pre-heated air in combustion processes are reported to be the main factors of influence on NOₓ formation. For hydrocarbon combustion, the most relevant mechanism of NOₓ production is categorized into prompt NOₓ (Fenimore NOₓ), these being
observed in rich premixed or diffusion flames with lower flame temperatures. Low flame temperatures would be provided either for rich or lean premixed combustion \[1,2,4\]. Obviously, due to fuel-saving strategies, latter approach is rather preferred. Nonetheless, lean mixtures can induce flow instabilities, flame detachment and possible extinction. Therefore, it is a condition which is often avoided.

Instantaneous water heaters can have a fully pre mixed flame in which the air and fuel are delivered to a plenum just upstream of the burner. In this way the appliance can operate in optimal conditions for a wide power band and, as a consequence, the emissions are usually very low. On the other spectrum of the market, atmospheric boilers do not have a ventilator to supply the combustion air at the correct rate and rely upon the air drafted by natural convection to complete the fuel oxidation. Some of the combustion air is pre mixed in an open nozzle that entrains the ambient air by the jet momentum. The remaining air is drafted as a secondary supply by convection. Emissions are typically much higher.

Zhang et al. \[2\] and Liu et al. \[3\] conceived an eco-friendly domestic burner element design that produces less than 40 ppm concentration of NO\(_x\) emission for a heat load of 22kW. This paper presents a slightly different design from the latter mentioned authors. In here most of the fuel is injected through a nozzle that supplies most of the thermal load on a lean mixture. The inherent instability of the flame is compensated by the existence of a secondary fuel injection that promotes a rich mixture in the vicinity of the main burner. The burner design reported by \[2,3\] has a single ejector per element whereas this report introduces two fuel jets per element. Both jets are discharged into separately mixing chambers which drive the air-fuel mixture to flame outlets. Mixture homogeneity and velocity profiles observed downstream in the burner element are crucial for combustion efficiency while reduction of pollutants formation is assured.

Both experimental and Computational Fluid Dynamics (CFD) techniques are challenging when considering the investigation on air-fuel mixture flow pattern. Although experimental methods are constrained for safety rules and cost, likewise CFD techniques also embrace several demanding issues. Spalart \[14\] considered CFD difficulties to consist of turbulence modelling, computing routines efficiency, numerical error minimisation and discretisation of both time and space. Although CFD simulations cannot stand for itself, requiring reliable experimental validation, they are widely used in pre-design and optimisation studies which enables project time and costs to diminish \[3,15\].

This paper describes the results obtained for CFD simulations of the air-fuel (propane and butane) mixture flow that comprises the domain between jets discharge plane and flame outlets on the top of the low NO\(_x\) element burner. Velocity profiles and distribution of excess air factor along flame outlets will be presented. For the present study in particular, which was conducted by numerical modelling using commercial software package FLUENT 15.0, we simultaneously dealt with jet spread and entrainment phenomena, multispecies transport, turbulence modelling and convergence criteria while addressing methodological and systematic procedures for results validation.

2 COMPUTATIONAL METHODOLOGY

2.1 Conservation equations

The flow is assumed to be in steady state, isothermal and incompressible. The latter
assumption is provided as a result of the fact that flow Mach number confirms to be less than 0.3 in every simulation conducted. The system of equations that describe the flow of a mixture of various fluids of a known composition.

\[
\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0 \tag{1}
\]

\[
0 = -\frac{\partial p}{\partial x} + \rho g_x + \frac{\partial\left(\mu \frac{\partial u}{\partial x} - \rho u^2\right)}{\partial x} + \frac{\partial\left(\mu \frac{\partial u}{\partial y} - \rho u v\right)}{\partial y} + \frac{\partial\left(\mu \frac{\partial u}{\partial z} - \rho u w\right)}{\partial z} \tag{2}
\]

\[
\nabla(\rho \cdot V \cdot Y_i) = -V \cdot J_i + S_i \tag{3}
\]

Equation 1 expresses the mass conservation principle. It should be noted that the flow velocity field \((\mathbf{V})\) is expressed as a function of \(u(x), v(y)\) and \(w(z)\), these representing each velocity component in tri-dimensional space. Equation 2 is the Reynolds averaged momentum (Navier-Stokes) conservation equation, depicted here for the \(xx\) component. Two additional equations in \(v(y)\) and \(w(z)\) were also considered. In the previous equations the variables are defined as density \((\rho)\), pressure \((p)\), gravity force \((g)\) and viscosity \((\mu)\). When, for instance, \(\bar{u}\) or \(u'\) notation appears, it refers to manipulated terms of velocity according to Reynolds-averaging technique [16,17]. Finally, Equation 3 addresses mass fraction conservation of the species considered to be present in the flow. The variables \(J_i\) and \(S_i\) represent the diffusive and convective terms, respectively and \(Y_i\) is local mass fraction for each species \(i\) [17].

2.2 Turbulence model selection

Although several approaches to turbulence modelling have been proposed in the course of the last century, none of the models conceived is reported to be able to predict accurately all the turbulent features for every condition [18]. In addition, the computational cost is a pressing issue for applications of complex shapes that renders more precise approaches (such as LES) impractical. The literature explains each model, though the wide scope of different methods' complexity, suits better some particular configurations. Considering this fact, a sturdy-scientific procedure when selecting the adequate turbulence model consists of applying a method that was previously validated for a configuration similar to the current one. Similar configurations are addressed if similar geometries, calculation domains and boundary conditions are involved [19,20].

Elattar et al. [21], in their investigation on CFD simulation of jet flames, evaluated the performance of various turbulence models. The authors concluded that the eddy-viscosity model \(k-\epsilon\) realizable and RSM were in best agreement with their both analytical solution and experimental data. However, the first mentioned method, based upon Reynolds-Averaged Navier-Stokes (RANS) equations, affords much less than RSM which is the reason why \(k-\epsilon\) Realizable was implemented in the current study. Several authors confirm \(k-\epsilon\) Realizable to has the more suitable performance when dealing with mixture and jet flows [17]. The eddy-viscosity \(k-\epsilon\), either standard or realizable model, uses the Boussinesq hypothesis to describe Reynolds stresses as a function of velocity gradients in the flow. Also transport equations for turbulence kinetic energy \((k)\) and turbulence dissipation rate \((\epsilon)\) are solved and when derived as proposed by Launder & Spalding (1974), in its standard formula, the model succeeds to predict jet spreading in a 2D domain. It does not occur the same way when performing a
circular axisymmetric jet flow simulation and the transport equations are reconceived to address this typical anomaly what originates $k–\varepsilon$ realizable [17].

2.3 Geometric model and boundary conditions

A geometric model was created based upon measurements made with a digital paquimeter on a real element burner previously provided. The configuration was conceived to comprise both nozzles involved and between-space upstream admission orifices, where atmospheric air entrains the fuel jet providing the partially pre-mixed combustion reactants. Both jet discharges were set 5 mm upstream admissions orifices, as proposed by [3], and have orifices of 0.75 mm in diameter. Geometric model and the most relevant boundary conditions implemented in FLUENT 15.0 can be seen in Figure 1.

Considering computational cost reduction, the implemented model only takes into account one half of the virtual configuration what can be easily overcome with a symmetry boundary condition applied to plane $z=0$ mm. The mass flow inlets were set for the domestic appliance with a heat load of 19 kW. It must be noticed the total thermal power of the burner is provided by an array of 12 elements such as that illustrated in Figure 1. Boundary (c), in Figure 1, was constrained by a pressure condition of 0 Pa since the fluid is discharged into the atmosphere. Likewise, despite not being graphically attended for simplifying purposes, the same pressure outlet condition was applied to the surfaces that comprise the half-cylindrical body where the entrainment of ambient air is expected to occur. The remaining boundaries which have not been specified yet were assumed to be solid walls, implying that no-slip conditions were applied. Similar boundary conditions are reported in studies conducted by other authors [2, 3, 15, 21, 22].

![Figure 1 - Geometric model and principal boundary conditions](image)
Turbulence conditions for mass flow inlets and entrainment region were specified upon nozzles’ turbulence intensity and hydraulic diameter. Turbulence intensity is calculated as in Equation 4 and was set as 6%.

\[
I = 0.16 \frac{Re}{\pi}
\]

(4)

Other turbulence conditions remained equal to the original FLUENT settings, which consists of 5% for turbulence intensity and 10% for turbulent viscosity ratio.

When performing the species transport simulation, for each boundary referred above, the user must specify \( n-1 \) mass fractions when considering \( n \) species to be present. For the mixture flow cases reported in this paper, a fuel mass fraction of 1 has been specified for the inlets while boundaries, with atmospheric pressure imposed, are specified in terms of air main species components. The air, dry, at 288.16 K, was assumed to be a fixed mixture of the main components: 0.232 for oxygen, \( 4.58 \times 10^{-4} \) for carbon dioxide, 0 for water. It makes up 1 as the software consider the remaining fraction to be nitrogen.

To better distinguish primary and secondary outlets, they were named as in Figure 2 according to their own symmetry planes. In Table 1, surface areas of admission orifices and outlets are recorded – it should be referred that both symmetric parts of the model were considered for surface areas calculation.

<table>
<thead>
<tr>
<th>Surface</th>
<th>Area [m²]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary admission orifice</td>
<td>1.44×10⁴</td>
</tr>
<tr>
<td>Secondary admission orifice</td>
<td>5×10³</td>
</tr>
<tr>
<td>Primary outlets</td>
<td>1.632×10⁴</td>
</tr>
<tr>
<td>Secondary outlet</td>
<td>1.28×10⁴</td>
</tr>
</tbody>
</table>

Figure 2 - Primary and secondary outlets
2.4 Numerical approach

FLUENT CFD software is based on the finite volume method. Amongst the various pressure-based solver approaches, segregated algorithm SIMPLE was selected for pressure-velocity coupling. In brief, it means that for each iteration, the governing equations are computed once at a time. The momentum equation is solved for a known pressure distribution and the resulting velocity field is used to update the pressure field through the continuity equation. The Reynolds-Averaged Navier-Stokes (RANS) equations were solved for every case tested in the present study and they were enough to perform preliminary tests in which the fluid was assumed to be air. When considering the air–fuel mixture modelling, species transport and energy models were set active. Numerical modelling has been succeeded in predicting multicomponent flows, although Padoin et al. [23] reported that it has not been presented yet any CFD investigation considering the whole phenomenological theory on mass transfer based on Maxwell–Steffan and Fick's generalised laws. The energy equation is activated if one needs to compute enthalpy transport since species diffusion must be taken into account [23, 24].

Every scheme of spatial discretization is imposed a second order upwind which means that quantities at cell faces are calculated based on a linear reconstruction approach regarding nodes' values. Second order upwind is accurate and more affordable scheme accepted by ASME on CFD procedures for scientific validation [25]. Relaxation factors used in the calculation routine was adopted to be 1 for body forces and turbulent viscosity, 0.8 for turbulent kinetic energy and turbulent dissipation rate, 0.7 for momentum. Enhanced wall treatment was used to describe the transition between boundaries imposed by the no-slip condition and turbulent flow. Using a wall function, it is avoided to load the software with a near-wall very refined mesh which would increase drastically the computation time [17]. Hybrid initialization method was selected to compute flow initial solution.

2.5 Grid generation and convergence criteria

Firstly, some tests were made in order to assess the extension of domain influenced by a free jet discharge. For this purpose, three meshes were used with, respectively 207,659; 753,682 and 2,104,678 elements and the Grid Convergence Index (GCI) method was applied to analyze the effect of grid assuming the maximum magnitude of velocity vector in the flow to be the test variable [26]. Grid Convergence Index was calculated to be 0.0036% which means that there was no improvement with mesh refinement. In the course of testing procedures, we concluded any entrainment effect is observed beyond 15 mm of the jet axis. As a result of this analysis, the cylindrical domain surrounding the injector orifice (Figure 1) comprising jet discharge and the entrance to the burner orifices extends no further than 15 mm from each nozzle's centre in a xy plane.

The GCI method is highly recommended by the CFD community, though it is very demanding on what computational and time costs are concerned [26]. The computational domain defined by the burner boundary was meshed as a single body what assures continuity of the calculation domain. However, the final geometric model consists of different blocks modelled separately that were subsequently assembled into a single unit. This method allows the user to customize the mesh parameters for each part of the model. The grid generated has 2 606 370 elements which means if it was simply consisted of small finite cubes, each cubic
element would have 0.2151 mm in size, considering the total volume of the geometric configuration. The computational mesh in the symmetry plane (z=0 mm) is displayed in Figure 3. The solution was stipulated for residuals to be less than $10^{-4}$ for every governing equation solved.

3 RESULTS AND DISCUSSION

3.1 Air jet

Fluids discharged from both nozzles were assumed to be air. This case is the simplest numerical modelling since species transport model is not activated. Air dynamic viscosity and density were fixed as $1.820 \times 10^{-5}$ kg/(m·s) and 1.204 kg/m³, respectively. These properties remain equal for the following cases presented. All results and boundary condition values imposed will be presented as the symmetry condition has not been applied, and the two symmetric bodies were considered instead. The pressure at the inlets was set for 38,700 Pa.

Mass flow inlet for both primary and second nozzles was $2.7830 \times 10^{-5}$ kg/s. This results in mean velocity jet discharges of 52.3 m/s with Reynolds number equal to 2,596. Table 2 reports the mass flow rate observed for primary and secondary outlets. The variable $m_{air}/m_{inj}$ represents the ratio between entrained air mass flow rate on the respective outlet and jet mass flow rate.
Table 2 - Results for entrainment with air jets

<table>
<thead>
<tr>
<th></th>
<th>Mass flow rate [kg/s]</th>
<th>$\frac{m_{\text{air}}}{m_{\text{inj}}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary outlets</td>
<td>$1.9644\times10^{-4}$</td>
<td>7.06</td>
</tr>
<tr>
<td>Secondary outlet</td>
<td>$5.0884\times10^{-5}$</td>
<td>1.83</td>
</tr>
</tbody>
</table>

3.2 Air-fuel mixture

In this case, the jet fluid was considered to be fuel and for that reason, CFD simulation is performed for an air-fuel mixture. Simulations were carried out with propane and butane jets, its properties being taken for 288.16 K. For both fuels, boundary conditions were adopted to satisfy operating conditions of a burner using propane with a heat load of 19 kW. Operating conditions for propane and butane are displayed in Table 3.

Table 3 - Operating conditions of a burner with a heat load of 19 kW

<table>
<thead>
<tr>
<th>Fuel</th>
<th>Nozzle diameter [mm]</th>
<th>Relative pressure [Pa]</th>
<th>Mass flow rate per element burner [kg/s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Propane (C$_3$H$_8$)</td>
<td>0.75</td>
<td>2,375</td>
<td>$3.4153\times10^{-5}$</td>
</tr>
<tr>
<td>Butane (C$<em>4$H$</em>{10}$)</td>
<td>0.72</td>
<td>1,475</td>
<td>$3.4684\times10^{-5}$</td>
</tr>
</tbody>
</table>

Regarding the mass flow rate of propane per element burner, it means that the simulations must fix a fuel mass flow inlet of $1.70765\times10^{-5}$ kg/s for each jet. Pressure inlets are set for 2,375 Pa, according to Table 3. Mean velocity jet discharges of 20.2 m/s and 15.7 m/s are observed for propane and butane, respectively. This results in the following Reynolds numbers on the discharge planes: 3,647 for propane jets and 4,141 for butane jets.

The velocity profiles of air-propane mixture on the outlets of the element burner can be seen in Figure 4. The primary outlets are named as in Figure 2. Considering the configuration of the burner element for the primary path, it seems that the pressure losses would increase as the mixture flow trajectory becomes longer. That appears to be the reason why in Figure 4 it can be seen the mean primary velocity profile decrease as the outlets goes from 4 to 1 (see the number reference in Figure 2).

However, the wedge-shaped mixing chamber improves the uniformity pressure distribution which added to the outlet partitions on the top of the element burner enables a better control on the outflow velocity profiles [2]. Of course, the velocity is zero near the partition walls as it was expected. For each primary outlet, the velocity grows in the negative $x$ direction as viscous forces introduced by no-slip condition are being dissipated. The secondary outlet mean velocity is very smaller since secondary mixing chamber is conceived to produce high pressure loss.

Figure 5 presents the propane mass fraction pattern for the highlighted cross-section domains. Total mixtures mass flow rate can be divided up into fuel and air which allowed to compute the entrainment mass flow rate. Consequently, the excess air factors in the outlets for both cases with propane and butane fuels were calculated and the results are reported in Table 4.
It should be noticed the coefficients of excess air factors are computed for the outlet plane of the element burner itself, that meaning the air entrained downstream which will make up the excess air actual present in combustion are not considered. The coefficient of excess air ($\lambda$) is calculated as in Equation 5, where $(m_{\text{air}}/m_{\text{fuel}})_{st}$ represents the stoichiometric air-fuel ratio which is 15.65 for propane and 15.5 for butane.

$$\lambda = \frac{m_{\text{air}}}{m_{\text{fuel}}} \left( \frac{m_{\text{air}}}{m_{\text{fuel}}}_{st} \right)$$

Although it appears that the air/fuel ratio is non uniform at the burner outlets, the actual variation is very small, below 0.1%. Therefore it is reasonable to assume that the burner promotes a homogeneous air/fuel mixture. The primary and secondary injections flow through different and separate mixing chambers, this being the reason why mixture richness differs from primary and secondary flame outlets. The results confirm what was expected and desired: venturi configuration in the primary path flow is specially designed to cause a suction effect that increases the amount of air entering the admission orifice [2].

The secondary mixtures are approximately 3 times richer than the primary ones assuring efficiency and stability of the combustion process. Generally, low NOx water heater appliances are designed to have an excess air factor of around 1.3. Considering total mass flow rates of air and fuel on the outlets, this means 82.3% and 84.19% of air (according to a mass flow basis) is still requested for ideal air-propane and air-butane combustion, respectively.
Figure 5 - Propane mass fraction distribution

Table 4 - Excess air ratio for air-propane and air-butane mixtures

<table>
<thead>
<tr>
<th></th>
<th>Propane</th>
<th>Butane</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary outlets</td>
<td>0.32</td>
<td>0.28</td>
</tr>
<tr>
<td>Secondary outlets</td>
<td>0.11</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Excess air factor distributions observed along the element burner outlets for propane and butane jet fuels can be seen in Figures 6 and 7, respectively.

Figure 6 - Excess air factor distribution for propane (C₃H₈) jet fuel
Regarding the results plotted in Figures 6 and 7, the maximum fluctuations towards the mean value of excess air factor in each cases were computed to be the following, respectively for primary and secondary outlets: 0.149% and 0.009% for propane fuel; 0.222% and 0.051% for butane fuel. It must be concluded that the range of excess air factor distribution is highly narrow either for both fuels and primary or secondary outlets, though secondary mixture is better distributed.

![Figure 7 – Coefficient of excess air factor distribution for butane \((\text{C}_4\text{H}_{10})\) jet fuel](image)

The difference between both fuels is due to the differences in the momentum of the fuel jet released into the atmosphere that entrains air at varying rates.

4 CONCLUSIONS

FLUENT 15.0 commercial software package was employed to conduct a CFD investigation on the air-fuel mixture velocity profiles and excess air factor distribution observed in the flame outlets of a low NO\(_x\) water heater element burner.

This study concluded that the geometrical configuration of the element burner provides a highly homogeneous distribution of the coefficient of excess air along the flame outlets. It was also concluded that for an atmospheric water heat appliance consisting of element burners as the presented one, only a preliminary mixture of the combustion reactants occurs as a consequence of entrainment caused by jet fuel discharges – this implies that most of the air involved in combustion is entrained downstream of element burner outlets.

CFD simulations results need to be validated by experimental data which consists on the remaining main question that is currently being addressed.

REFERENCES


COMPUTATIONAL FLUID DYNAMIC ANALYSIS OF OLIVE OIL IN DIFFERENT INDUCTION POTS

J.P.KASTILLO1*, J.MARTÍNEZ1†, A.J.RIOFRIO1, S.P.VILLACIS1, M.A. OROZCO1

1 Instituto Nacional del Eficiencia Energética y Energías Renovables (INER)
Avda. 6 de Diciembre N33-32 e Ignacio Bossano Edificio Torre Bossano - Código Postal: 170511
Quito – ECUADOR.
juan.kastillo@iner.gob.ec
javier.martinez@iner.gob.ec
augusto.riofrio@iner.gob.ec
stefany.villacis@iner.gob.ec
marco.orozco@iner.gob.ec

Key words: Induction cookware, induction cooking, heat transfer analysis, cast iron, thermography, COMSOL Multiphysics, Computational fluid dynamic.

Abstract.
Induction cookers are popular in modern families around the world, because they present several advantages in cooking over gas and electrical resistance stoves such as higher energy efficiency, higher heating rates and safety. However the appropriate performance of these devices for induction heating (IH) depends directly on the coupled system cooker-cookware. Differences in the energy efficiency and temperature distribution during heating and cooling are related to the configuration of the pot and the different materials of the pot. In case of cooking processes, the natural convection of the different fluids may affect to the microbiology, safety, food quality and cooking times. This study aims to perform a convection and temperature distribution of olive oil analysis of three different induction cookware configurations of aluminum, enameled iron and stainless steel body cookware. The experimental data was analyzed using computational fluid dynamic (CFD) software in order to understand the thermal processes in the different pots. These results were compared with the results obtained by a thermographic camera. Differences in convection flow distribution during IH were observed on the measurements and CFD analysis and these are related with the materials configuration of the pot bottom.
1 INTRODUCTION

The induction heating (IH) technique has been developed since 1970, within American, European and Asian industries, they are recently becoming popular for heating food in households [1], [2], [3], [4]. The induction cooker directly heats the cookware bottom through an electromagnetic field, instead of doing it through radiant heating as in electrical resistance cookers or in liquefied petroleum gas (LPG) based cookers. These kinds of cookers use an alternating magnetic field that magnetizes the ferromagnetic material of the cookware in both directions. This process has less energy losses, the material magnetically agitates, the absorbed energy releases as energy, heating the cookware.

The cookware must be made of a ferromagnetic material at least in its bottom. Induction cooking presents benefits regarding other kind of cookers relate with a better energy efficiency than conventional LPG based cookers and resistance electric cookers due to the reduction of heating losses with the environment by the induced magnetic field in cookware. Besides it is able to get a higher speed of heating for the cookware bottom ferromagnetic material capacity to attract magnetic fields, as a result the cookware directly heats and not the surface. They are safer because they do not produce fire, there are less possibilities of getting burn. The vitroceramic allows to perform an easily cleaning of the cooker surface and they have digital control so it becomes easy to manage it and there is not LPG combustion near to the user. [1], [6], [7].

The Ecuador Government is looking forward to replace 3 million LPG based cookers with induction cookers, besides the introduction of cookware sets consisting of three pots with different diameters and a saucepan, being a world pioneer campaign called “efficient cooking plan”. Computational fluid dynamics (CFD) models have been applied since long time ago to different processing industries, such as the aerospace, automotive and nuclear industries and in the last two decade have been applied to food processing applications [8].

In the view of food engineering, CFD can be defined to be a numerical methodology that has been used for solution of heat, mass and momentum (fluid flow) transfer equations simultaneously with the given boundary conditions in a given computational domain [9]. Many food processing operations such as chilling, drying, baking, mixing, freezing, cooking, pasteurization and sterilization rely on fluid flow [10]. Some of the most relevant applications of CFD in the food industry were summarized in [8], [9].

The phenomenon of natural convection in fluid-filled has been investigated with the help of either numerical analysis using in-house or open source of commercial CFD software sometimes coupled with experimental techniques of flow visualization and temperature measurement [8]. Some of the CFD applications to simulate convection flow are mentioned in [8].

In relation to IH modeling, [8] investigated the IH process using a non-contact measurement method based on thermography. They demonstrated that thermo-graphic
measurements are useful tools for observing temperature distributions of work-pieces during IH. Evaluated a non-linear and transient magneto-thermal coupled problem with a moving coil which is considered as moving heat source, by a multi analytic software [8].

A study of the heating process of the pots bottom in an induction cooker with a pot filled with water has been made. [8]. In this sense, this research aimed to perform a new analysis of the heating process of oil in the bottom and in the sides of three types of pots, with different configurations, speaking about the their body and bottom material, in order to know the temperature distribution, speed of heating and cooling, and the heat flux of pots during heating, in vitrified iron, stainless steel and aluminum body with stainless steel bottom pots, in order to determine the most suitable pots for induction. A physical model has been studied using the multi analytic software COMSOL Multiphysics© and it has been validated with data obtained from a thermographic camera, through laboratory tests.

2 MATERIALS AND EXPERIMENTAL METHOD

2.1 Procedure

The procedure of the performed tests consists of pots heating with and without water, within laboratory conditions. The pots were located on an induction cooker, model Povos/Pentium C21-PH12T, it has only one induction zone with a nominal power of 1 000 W and a voltage of 110 V. The thermographic camera, model FLUKE Ti125 was used, it has a temperature scale from 60 °C to 385 °C. The thermographic camera was set perpendicular to the pot with a separation of 1,20 m as is shown in Fig. 1, this procedure was followed for the pot bottom and also for its sides. After that, the cooker was turned on while the thermographic camera was recording a video and saving in its memory the outstanding temperatures for this study. It was able to get the thermographies of the bottoms and laterals at different instants of time and this has allowed to get the temperature distribution for several points of the pot bottom and laterals. It is important to notice that these tests were performed in the UMCO cookware factory, located in the south of Quito, which is about 2800 meters above sea level, and has an atmospheric pressure of 728.4 hPa, which is the reason of the water boiling at 92 °C.

![Fig. 1 Scheme of the procedure of recording measurements with the thermographic camera](image-url)
Table 1 Characteristics of tested pots

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Enameled iron</td>
<td>Enameled iron</td>
<td>17.00</td>
<td>20.00</td>
<td>0.7</td>
<td>1.0</td>
<td>53.3</td>
<td>7.18</td>
<td>500</td>
<td>468.00</td>
</tr>
<tr>
<td>2</td>
<td>AISI 304 Stainless steel</td>
<td>AISI 430 Stainless steel</td>
<td>17.00</td>
<td>20.00</td>
<td>0.5</td>
<td>1.8</td>
<td>24.9</td>
<td>7.80</td>
<td>505</td>
<td>900.50</td>
</tr>
<tr>
<td>3</td>
<td>Aluminum</td>
<td>AISI 430 Stainless steel, Acero Inoxidable</td>
<td>17.00</td>
<td>20.00</td>
<td>2</td>
<td>2.5</td>
<td>209</td>
<td>2.70</td>
<td>909</td>
<td>744.00</td>
</tr>
</tbody>
</table>

Table 1 shows the main characteristics of the pots that were used in the tests, it has been chosen different material pots suitable for induction. The first configuration of cookware was tested in a pot made of enameled iron in its body and in its bottom; then the second cookware configuration was tested, it consists of pot made of AISI 304 stainless steel in its bottom and in its base a three layer material made of AISI 430 stainless steel, aluminium and iron, all of them weld together to its body and covered by an AISI 430 stainless steel layer. The last cookware configuration was tested, in a pot made of aluminium in its body and AISI 430 stainless steel in its bottom.

Fig. 2 Pot bottom and body

a= Diameter of pot bottom.

b= Height of pot body.

Fig. 2 shows a diagram of the points that have been taken to perform the temperature analysis from the thermographic images within the temperature analysis during heating and cooling, central and extreme points were chosen. On the other hand, for heat transfer analysis, it was chosen a line with 112 points. The values of temperatures were taken every five seconds; in the case of pots without water, measurements were taken until saturation temperature was reached, or until the temperature sensor of the cooker reaches its limit and it
turns off. During the test of pot with water, temperature values were measured every ten minutes.

Once that the heating test was performed, the cooker was turned off and the temperature during pot cooling was measured, until the pot central zone reached a temperature of 50 °C in the case of pots without water, and during ten minutes in the case of pots with water. The analyzed points were used to plot graphs of temperature against time.

The obtained temperature data was used to make a computational fluid dynamics (CFD) analysis, in order to get results related with convective heat flux, through the modulation and simulation with the software COMSOL Multiphysics©.

**Nomenclature**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>area (m²)</td>
</tr>
<tr>
<td>Cₚ</td>
<td>specific heat (J/kg °C)</td>
</tr>
<tr>
<td>g</td>
<td>gravitational acceleration (m/s²)</td>
</tr>
<tr>
<td>h</td>
<td>heat transfer coefficient (W/m °C)</td>
</tr>
<tr>
<td>k</td>
<td>thermal conductivity (W/m °C)</td>
</tr>
<tr>
<td>p</td>
<td>static pressure (N/m²)</td>
</tr>
<tr>
<td>q</td>
<td>heat flux (W/m²)</td>
</tr>
<tr>
<td>t</td>
<td>time (s)</td>
</tr>
<tr>
<td>T</td>
<td>temperature (°C)</td>
</tr>
<tr>
<td>v</td>
<td>velocity (m/s)</td>
</tr>
<tr>
<td>V</td>
<td>electric scalar potential (V)</td>
</tr>
<tr>
<td>A</td>
<td>magnetic vector potential (V s/m)</td>
</tr>
<tr>
<td>B</td>
<td>magnetic induction (N s/C m)</td>
</tr>
<tr>
<td>⃗E</td>
<td>electric flux density (C/m²)</td>
</tr>
<tr>
<td>⃗H</td>
<td>magnetic field (Am)</td>
</tr>
<tr>
<td>⃗J</td>
<td>electric current density (A/m²)</td>
</tr>
<tr>
<td>β</td>
<td>thermal expansion coefficient (1/K)</td>
</tr>
<tr>
<td>ε</td>
<td>dielectric constant (F/m)</td>
</tr>
<tr>
<td>ρ</td>
<td>density (kg/m³)</td>
</tr>
<tr>
<td>µ</td>
<td>apparent viscosity (Pa s)</td>
</tr>
<tr>
<td>µₘ</td>
<td>magnetic permeability (N/A²)</td>
</tr>
<tr>
<td>r</td>
<td>radial direction of the pan</td>
</tr>
<tr>
<td>θ</td>
<td>angular direction of the pan</td>
</tr>
<tr>
<td>z</td>
<td>vertical direction of the pan</td>
</tr>
<tr>
<td>ref</td>
<td>reference</td>
</tr>
<tr>
<td>w</td>
<td>wall</td>
</tr>
</tbody>
</table>

**2.2 Assumptions**

To simplify the problem, the following assumptions were made:

1. Axi-symmetry: The shape of the pot was assumed to be a cylinder. The convection of fluid may show axial symmetry in the pot because the heating condition is axi-symmetrical.

2. Heat generation due to viscous dissipation is negligible.

3. Boussinesq approximation was applied for the buoyancy.

4. The assumption of no-slip condition at the inside wall of the pot is valid.

5. Convection in the pot was a laminar flow, because it was a natural convection.

6. The gravity force of 9.8 m/s² was added axially in the pot.

The thermophysical properties of the oil were considered temperature dependence, and their values change during CFD simulation.

**2.3 Model equations**

The partial differential equations governing natural convection motion in a cylinder space are the Navier–Stokes equations in cylindrical coordinates [8] as shown below:
(1) Continuity equation
\[ \frac{1}{r} \frac{\partial}{\partial r} (rp v_r) + \frac{1}{r} \frac{\partial}{\partial \theta} (pv_\theta) + \frac{\partial}{\partial z} (pv_z) = 0 \] (1)

(2) Energy conservation equation
\[ \frac{\partial T}{\partial t} + v_r \frac{\partial T}{\partial r} + \frac{v_\theta}{r} \frac{\partial T}{\partial \theta} + v_z \frac{\partial T}{\partial z} = k \left[ \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial T}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 T}{\partial \theta^2} + \frac{\partial^2 T}{\partial z^2} \right] \] (2)

In order to simplify the analysis, the buoyancy force caused by density variation with temperature was governed by the Boussinesq approximation shown in Eq. (3), which was used in the body force term of the momentum equation in the radial direction.

(3) Momentum equation
- In the radial direction (r)
\[ r \left( \frac{\partial v_r}{\partial t} + v_r \frac{\partial v_r}{\partial r} + \frac{v_\theta}{r} \frac{\partial v_r}{\partial \theta} - \frac{v_\theta^2}{r} + v_z \frac{\partial v_r}{\partial z} \right) = -\frac{\partial p}{\partial r} + \mu \left[ \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial v_r}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 v_r}{\partial \theta^2} + \frac{\partial^2 v_r}{\partial z^2} \right] \] (3)

where \( T_{ref} \) and \( \rho_{ref} \) are the reference temperature and density, respectively.
- In the vertical direction (z)
\[ \rho \left( \frac{\partial v_z}{\partial t} + v_r \frac{\partial v_z}{\partial r} + \frac{v_\theta}{r} \frac{\partial v_z}{\partial \theta} + v_z \frac{\partial v_z}{\partial z} \right) = -\frac{\partial p}{\partial z} + \mu \left[ \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial v_z}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 v_z}{\partial \theta^2} + \frac{\partial^2 v_z}{\partial z^2} \right] \] (4)

- In the angular direction (\( \theta \))
\[ \rho \left( \frac{\partial v_\theta}{\partial t} + v_r \frac{\partial v_\theta}{\partial r} + \frac{v_\theta}{r} \frac{\partial v_\theta}{\partial \theta} + v_r v_\theta \frac{\partial v_\theta}{\partial z} \right) = -\frac{1}{r} \frac{\partial p}{\partial \theta} + \mu \left[ \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial v_\theta}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 v_\theta}{\partial \theta^2} + \frac{2}{r^2} \frac{\partial v_\theta}{\partial \theta} + \frac{\partial^2 v_\theta}{\partial z^2} \right] \] (5)

2.3 Analysis of electromagnetic field

The electromagnetic model is classically based on the set of Maxwell equation. This system is based on the four following equations:
- Magnetic flux
\[ \nabla \cdot B = 0 \] (8)
- Maxwell – Gauss
\[ \nabla \cdot E = 0 \] (9)
- Maxwell – Faraday
\[ \nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \] (10)
• Maxwell – Ampere

\[ \vec{\nabla} \times \vec{H} = \vec{j} + \frac{\partial \vec{\psi}}{\partial t} \]  

(11)

2.4 Solution strategy

The software used in our simulations of the convection flow of oil in the pans during IH was COMSOL which is based on the finite volume method.

To shorten the computation time, calculations were done for only half of the pan. However, when convection occurred in two kinds of fluids, such as oil and air, in a single model, the calculation is complex, and took a long time. It is assumed in such cases that there is no convection but only conduction in the air layer, since the target fluid in this analysis is oil.

CFD was customized by COMSOL programming to introduce the heat generation into the CFD calculation. By a so-called skin effect, the high-frequency current or the electromagnetic field is limited to the surface of the conductor. We assumed the measurements of the different pots. Fig. 3 shows the heat generation distribution of the bottom of the pan as determined by electromagnetic field analysis (EFA). This figure shows the distribution in the radial direction; where 0, on the horizontal axis, is the center of the pan. Fig. 3 also shows the temperature distribution of the bottom of the three pots determined by the radiation thermometer. The heat generation distribution was founded similar to the temperature distribution. The heat generation distribution was divided as shown in Fig. 3, and their values were used for the CFD analysis.

![Fig. 3 Heat generation distribution of the bottom of the pan at several positions, determined by electromagnetic field analysis, with its body made of a) enameled iron, b) stainless steel, c) aluminum.](image)

3 RESULTS

3.1 Observation of heating location with radiation thermometer

After the empty pan was heated for 15 s, the temperature distribution of the pan collected by the radiation thermometer is shown in Fig. 4 a), b), c) (top view) and Fig. 4 d), e), f) (side view). In the enameled iron, the region of the pan just above the coil was heated in a doughnut
shape, and this heated part rose to a high temperature. The temperature on the sides of the pan was kept in low. Results are in agreement to [8] and [23], they reported that during IH a single coil heater produce a non-uniform temperature distribution over the cross and depth sections of the heated pan, because the skin effects. In case stainless steel and aluminium pot the distribution of temperatures are similar in the base, [21] explained that in the stationary induction, heat treating the contact area between the coils and work-piece, and since this area is a small part of the whole area, a lot of power is required to extend the heat to the whole area from the small contact area. For this reason often the use of ferrite or laminated loading materials is required to influence the magnetic field concentration in given areas thereby to refine the heat pattern produced. Note these results are with an empty pan and the fluid change the distribution of temperatures above the base.

![Fig. 4 Temperature distribution of the pan collected by the radiation thermometer heated by IH after 15 s. Top view and side view for a), d) enameled iron, b), e) stainless steel, c), f) aluminium.](image)

### 3.2 Heat flux measurement

The temperature distribution of the bottom of the pan and the heat flux in the pan were compared. The results under IH and GRH are shown in Fig. 5 a), b) and c), respectively. These figures revealed a correlation between the heat flux and the temperature of the pan, even a few measured points of heat flux were considered, just to confirm the accuracy of the
distribution showed in Fig. 4 and validate by this way its use as boundary conditions in case of IH.

Results in Fig. 5 a) confirmed those showed in Fig. 4 a) in relation to the doughnut shape of the temperature profile of the bottom of the pan under IH when it has been increase the heating time. A similar behavior appears in Fig. 5 b) and c) at 45 s of the heating time.

![Heat flux viewed from the pots bottom with its body made of (a) enameled iron, (b) stainless steel, (c) aluminium, during heating process at 15 s, 30 s and 50 s.](image)

**Fig. 5** Heat flux viewed from the pots bottom with its body made of a) enameled iron, b) stainless steel, c) aluminium, during heating process at 15 s, 30 s and 50 s.

### 3.2 Comparison of measured and calculated temperature change of the oil in the pot

The measured and calculated local temperature change in the pots was compared as shown in Fig. 6 and for a stainless steel pot. The location of the measured position is shown in Fig. 2. The vertical axis shows the difference from the initial temperature. From these figures it can be seen that the rise speed of the temperature rate is almost same according to the location between the calculated and the measured values after 125 s, however, a slightly deviation was observed before this time. This irregular behavior could explain the higher differences between the calculated and measured temperature change of the fluid in the pan presented in Fig. 6. Further the simplification of the model in terms of consider only laminar fluid may affect the modeled convection at high heating temperature observed in the present study, as well as the effect of the computational grid of the model used in simulation that showed a non-uniform construction being uniform in axial direction but non-uniform in radial direction.
3.4. Visualization of temperature distribution in the pot

The temperature change of the fluid including the wall of the pot can be shown as an animation in COMSOL. Fig. 7 shows the temperature distribution from the animation at 300 s. The right extremity of the figure is the side of the pot, and the left extremity of the figure is the center of the pot. The temperature at the bottom of the pot increased as soon as the cookware was heated. Oil was heated along with the heating of the bottom, and it can be seen that a part of the heated oil was rising until reach the top water surface after 300 s (Fig. 7 a), b) and c)). The temperatures at the side wall of the pan did not really increase.

Fig. 8 shows the calculation results for the resultant velocity convection flow in the pot heated by IH at 10 s, 60 s, and 300 s. The right extremity of the figure is the side of the cookware, and the left extremity of the figure is the center of the pot. The length of the arrow shows the magnitude of convection and the direction of the arrow shows the direction of convection. While the convection had scarcely occurred at 10 s of heating, the convection had risen from the inside of the pan and moved outside at 60 s. The tendency of convection from the inside to the outside was strengthened at 300 s of heating.

In other words, the oil adjacent to the bottom surface above the coils received a higher induction heat in enameled iron and stainless steel but rose in the border for aluminium pots. As the liquid is heated, it expands and thus gets lighter. Liquid away from the coils position stays at a much lower temperature. The buoyancy force created by the change in liquid density due to temperature variation produces an upward flow, near the coils position (from the bottom of the pot to the top). The hot liquid going up is deflected at the top and then travels radially towards the side of the pot. Being heavier, the oil near to the sides moves downwards and then towards the coils position against.
Fig. 7 a) Visualization of temperature distribution in the pan heated at 300 s for a) enameled iron, b) stainless steel, c) aluminum.

Fig. 8 Calculated convection flow in the pots of a) enameled iron, b) stainless steel and aluminum after 10, 60 and 300 s

4. Conclusion
Convection in a pots heated by IH was analyzed using CFD. The heat transfer was set so that the side of the pot might become a cooling surface.
The rate of increase in oil temperature, which was calculated by CFD, corresponded to the measurement result. The simulation of convection within the pot showed almost the same tendency as the movement of the tracer particles in flow visualization. The direction and resultant velocity of convection differed between the heating appliances, especially at the initial stage of the heat, however at high heating temperature the resultant velocity become similar under both heating types. Thus, the velocity and direction of convection flow in the pan could be estimated by CFD analysis.

Acknowledgements

The authors of this present research acknowledge to the Secretaria Nacional de Planificación y Desarrollo (SENPLADES) for financing the execution of the present research. This work was sponsored by the Prometeo project of the Secretaria de Educación Superior, Ciencia, Tecnología e Innovación (SENECYT) held in the Republic of Ecuador. The information necessary to complete this work was given by the Ministerio de Electricidad y Energía Renovable (MEER) of Ecuador.

References


NUMERICAL STUDIES OF PARAGUAYAN CHACO NATURAL GAS IN A COUNTERFLOW DIFFUSION FLAME

HERNANDO MALDONADO1, DARIO ALVISO, MIGUEL MENDIETA, JUAN C. ROLON

1Laboratorio de Mecánica y Energía, Facultad de Ingeniería, Universidad Nacional de Asunción 2160 Campus de la UNA, San Lorenzo, Paraguay hernandomaldonado90@gmail.com

Key words: Natural Gas, Counterflow, Diffusion Flames, Kinetic Modeling

Abstract. Combustion is a chemical phenomenon in which a multitude of elementary chemical reactions take place, resulting in the overall process of fuel oxidation. Natural gas is a fuel that has been explored for a few decades and extracted for a few years in the region of Paraguayan Chaco, near Bolivia border. Currently, natural gas is not very important in Paraguay’s energy matrix, however it could be in the near future if higher volumes are extracted and transported to the most populated cities, specially to Asunción. In order to improve combustion performance, an understanding of its fundamental properties and the combustion pathways is required. This study presents new data for Paraguayan Chaco natural gas combustion in a laminar counterflow diffusion flame configuration at atmospheric pressure. The visible chemiluminescence of the excited radicals CH* and C2* is employed experimentally. Numerically, a 1D simulation was carried out based on the specific natural gas chemical composition and using a standard kinetic mechanism, to which we add CH* and C2* reactions. Resulting from simulation the typical flame structure is presented and a tentative validation of the model is realized comparing experimental and numerical CH* and C2* radicals profiles.

1 INTRODUCTION

Mastering energy is a key factor for human to preserve its economic growth and maintain his life standard. According to the International Energy Agency (IEA) report and Shahid et al. [1], the world will increase by 40 % its energy requirements in 2035.

Taking into account all the primary energy produced around the world, combustion of liquid, gas and solid fuels corresponds to about 85% of the total, and as these resources are limited, combustion efficiency is very important, and also due to climate changes,
lower emissions to the atmosphere is one of the main requirements nowadays in most combustion systems and a challenge for research.

Combustion is a chemical phenomenon in which a multitude of elementary chemical reactions take place, resulting in the overall process of fuel oxidation. During these elementary chemical reactions, the chemical species are transformed into other species. These reaction intermediates may absorb part of the reaction energy placing them in an excited state. A possible way of de-excitation is the spontaneous emission of photons from the populated state of the specific species, called chemiluminescence. Chemiluminescence is the emission of light by molecules returning to their ground state after being excited chemically, rather than thermally. The emission shows a weak dependence on temperature and a greater dependence on chemical composition. One way to understand these complex chemical processes, is to follow the evolution of chemiluminescence of specific species produced in the reaction processes.

In order to acquire a better understanding of combustion fundamental phenomena, it is preferable to study flames configurations with flows relatively simple that are easy to control.

The counterflow laminar flames meet these criteria: the flow is well known, it can be modeled from a theoretical point of view and experimentally it can be controlled in a relatively easy manner.

The interest of these stretched flames also lies in the fact that they act as the main element to describe some turbulent combustion models.

The primary combustion zone of hydrocarbon flames is distinct from the other parts of the flame by its emission of mostly visible bands, namely $CH$ bands at 431 nm ($A^2 \Delta \rightarrow X^2 \Pi$) and $C_2$ band at 516 nm ($d^3 \Pi_g \rightarrow a^3 \Pi_u$) (Swan band).

A convenient way to experimentally study the flame behavior is to analyze the space and time resolved emission of excited species $CH^*$ and $C_2^*$, which are not at equilibrium under flame conditions. Indeed, these radicals are naturally present in the combustion zone and permit to determine important macroscopic properties such as flame location, flame speed, equivalence ratio and heat release rate fluctuations.

Natural gas is a fuel that has been explored for a few decades and extracted for a few years in the region of Paraguayan Chaco, near Bolivia border. Currently, natural gas is not very important in Paraguay’s energy matrix, however it could be in the near future if higher volumes are extracted and transported to the most populated cities, specially to Asunción.

This work aims to study the combustion of Paraguayan Chaco natural gas, from an experimental and numerical point of view, for which the counterflow diffusion (non-premixed) flame configuration at atmospheric pressure and constant injection temperature was chosen. We intend to study two operating flame conditions (equivalence ratio and strain rate), for which the existing experimental and computational equipments of the Laboratory of Mechanical and Energy of FIUNA was used. The flame structure of natural gas diffusion flame will be analized. $CH^*$ and $C_2^*$ experimental and numerical
profiles will be compared in order to validate the chosen chemical mechanism.

2 EXPERIMENTAL SETUP

2.1 Natural gas chemical composition

Paraguayan Chaco natural gas chemical composition was analyzed using Gas Chromatography and Mass Spectrometry (GC-MS) technique. This information is needed to estimate the equivalence ratio of the natural gas flame, and also it will be used as input to carry the simulations.

The estimation of natural gas is presented in Table 1. In comparison to the literature values, it corresponds to the classical composition of this fuel, where the major component is methane, with some minor components, as shown in Table 1. 'Others' (with a mole fraction x 100 of 0.987) refers to species not present in the chosen kinetic scheme. This point will be detailed in Section 3.2.

<table>
<thead>
<tr>
<th>Paraguayan Chaco natural gas composition</th>
<th>Mole fraction x 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nitrogen</td>
<td>3.095</td>
</tr>
<tr>
<td>Carbon dioxide</td>
<td>0.030</td>
</tr>
<tr>
<td>Methane</td>
<td>94.277</td>
</tr>
<tr>
<td>Ethane</td>
<td>1.611</td>
</tr>
<tr>
<td>Others</td>
<td>0.987</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 1: Composition of Paraguayan Chaco natural gas

2.2 Counterflow Burner Device

The experiments were carried out in a counterflow burner. The setup used will be described briefly. More information of the burner can be found in [2, 3].

The burner consists of two opposed axisymmetric convergent nozzles of 20 mm inner diameter (see Figure 1A) A flow of air is injected at ambient temperature through the upper side of the burner, while a mixture of natural gas and nitrogen is transported from the lower side also at ambient temperature. Nitrogen is injected in order to compensate the momentum of the air injected in the upper side. A laminar diffusion flame front appears between the nozzles. The distance between the two nozzle exits was kept constant to 20 mm. Each nozzle contains two layers of honeycomb to reduce turbulent structures in the plenum chamber of the nozzle. Each nozzle is surrounded by a coaxial nozzle fed by nitrogen, which isolates the reaction zone from ambient perturbations that could disturb the measurements.
In order to limit the thermal effect of the burnt gases circulating in the upper zone of the system, the upper burner has a thermal screen, as well as an isolator (glass wool) with plaster to optimize the temperature uniformity, as one can see in Figure 1A.

Two natural gas flame operation conditions have been chosen, considering the equivalence ratio and the strain rate: Flame I, with an equivalence ratio of $\phi = 4$ and strain rate $\epsilon = 134 \, s^{-1}$, and Flame II, with an equivalence ratio is $\phi = 3.5$ and strain rate $\epsilon = 136 \, s^{-1}$. The injection temperature was kept constant throughout the study in both nozzle exits (Injection Temperature = 300 K).

Strain rates ($dV/dz$) are estimated using the following equation:

$$\varepsilon = \frac{2V_u}{H} \left( 1 + \frac{V_l}{V_u} \sqrt{\frac{\rho_l}{\rho_u}} \right)$$

(1)

Where $V_u, V_l, \rho_u, \rho_l$ are the upper and lower nozzle exits velocities and densities, respectively. H is the distance between the two nozzle exits.

Figure 1: A) Counterflow burner device, B) Visualization set up
All gaseous mass flow rates are controlled by the mass flow meters with an accuracy of ±1.5% and a repeatability of ±0.5% of the full scale. The overall uncertainty of strain rates and natural gas flames equivalence ratios can therefore be estimated at ±0.78% and ±1.52%, respectively.

2.3 Visualization

Images of $CH^*$ and $C_2^*$ emission have been recorded using a 576 × 384 pixels ICCD camera (Princeton Instruments) equipped with UV-lens (Nikkor 105 mm focal length) (see Figure 1B).

Narrow-band interference filters were interposed along the optical path for capturing the $CH^*$ and $C_2^*$ emission. The filter used for $CH^*$ has 60% transmission and a 10 − nm wide bandpass centered around 430 nm. The $C_2^*$ filter is centered at 515 nm, 30 − nm bandpass and 75% transmission in the maximum.

Figure 2A presents a typical natural gas flame front obtained from $CH^*$ visualization. This figure shows that the flame front is slightly convex due to gravity effects. The line of sight creates an integrated signal along its trajectory and increases the measured mole fraction. As the flame front is axi-symmetrical, Abel integral inversion is used to eliminate the effects of the integration and get the trace of the flame front in the symmetry plane of the burner as done in [4, 5] (Fig. 2B, Fig. 3). It should be noted that the Abel inversion has shifted slightly the location of the maximum $CH^*$ emission intensity. Axial profiles of both $CH^*$ and $C_2^*$ are normalized, because absolute measurements were not made. Therefore, the comparison with numerical profiles will be done considering the position of both $CH^*$ and $C_2^*$ radicals, and not the concentration.

![Averaged image](image1)

**Figure 2**: A) Averaged image of Flame I $CH^*$ emission, B) Abel Inversion of the averaged image of Flame I $CH^*$ emission

The resolution of the CCD camera along the flame axis is ± 77 μm. Then, due to the slight flames oscillation, the flame position was estimated with an accuracy of ± 110 μm.
3 NUMERICAL SIMULATION

3.1 Governing equations

We consider an axisymmetric counterflow configuration shown in Figure 4. Air is injected from the left side whereas a mixture of natural gas and nitrogen is injected from the right side. We model our system using similarity approach by searching for similar solutions of gaseous flow equations in the vicinity of the central axis [7]. These similar solutions have the form: gas density \( \rho = \rho(z) \), gas radial velocity \( u = r U(z) \), gas axial velocity \( v = v(z) \), gas temperature \( T = T(z) \) and species mass fractions \( Y_k = Y_k(z) \), \( k = 1, \ldots, N_{sp} \) (\( N_{sp} \) is the number of species).

By assuming a pressure gradient in the radial direction so that \(-\frac{1}{r} \frac{\partial p}{\partial r} = J\) is constant along the \( z \) axis, the gaseous phase are described by the following balance equations [7]:

\[
2 \rho U + \frac{\partial \rho v}{\partial z} = 0, \tag{2}
\]

\[
\rho U^2 + \rho v \frac{\partial U}{\partial z} = \frac{\partial}{\partial z} \left( \mu \frac{\partial U}{\partial z} \right) + J, \tag{3}
\]

\[
\rho v c_p \frac{\partial T}{\partial z} = \frac{\partial}{\partial z} \left( \lambda \frac{\partial T}{\partial z} \right) - \sum_{k=1}^{K} h_k W_k \Omega_k - \left( \sum_{k=1}^{K} \rho Y_k V_{zk} c_{pk} \right) \frac{\partial T}{\partial z}, \tag{4}
\]
\[ \rho v \frac{\partial Y_k}{\partial z} = - \frac{\partial}{\partial z} \left( \rho Y_k V_k z \right) + W_k \Omega_k, \quad k = 1, \ldots, N_{sp} \quad (5) \]

In these equations \( \mu \) is the gaseous mixture viscosity. \( c_{pk} \) and \( c_p \) are the heat capacity at local constant pressure of species \( k \) and of the mixture respectively. \( h_k, W_k, \Omega_k \) are the specific enthalpy, the molar weight and the molar chemical production rate of the \( k^{th} \) species respectively, and \( V_{kz} \) is the diffusion velocity of the \( k^{th} \) species in the axial direction.

The above system of equations is completed by the ideal gas equation and by specifying the equation \( \frac{\partial J}{\partial z} = 0 \).

### 3.2 Numerical conditions

The boundary conditions of the system are given in Table 2.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Left ((z = -10))</th>
<th>Right ((z = 10))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gas temperature</td>
<td>( T_g = T_{gleft} )</td>
<td>( T_g = T_{gright} )</td>
</tr>
<tr>
<td>Species mass fractions ( Y_k )</td>
<td>( Y_k = Y_{kleft} )</td>
<td>( Y_k = Y_{kright} )</td>
</tr>
<tr>
<td>Gas axial velocity ( v_g )</td>
<td>( v_g = v_{gleft} )</td>
<td>( v_g = v_{gright} )</td>
</tr>
<tr>
<td>Gas radial velocity ( u_g )</td>
<td>( u_g = 0 )</td>
<td>( u_g = 0 )</td>
</tr>
</tbody>
</table>

**Table 2**: Boundary conditions

The set of equations is then replaced by a fully coupled set of discrete relations. The solution of this system is then based on a global adaptive nonlinear method using Newton
iterations (see [8, 9]). The grid is adapted to first and second order derivatives of all variables and the smallest grid size is 5 µm.

The kinetic modeling for natural gas oxidation in the counterflow flame was performed using the REGATH-1D-COUNTERFLOW code of the REGATH package with detailed thermochemical and transport properties developed at EM2C laboratory. The inputs to each simulation include a chemical kinetic reaction mechanism, a dataset of thermochemical properties and a dataset of transport properties.

A methane mechanism (Lindstedt et al. [10] including 29 species and 141 reactions) was chosen to carry the simulations. We have chosen this mechanism for the natural gas/air combustion due to the composition of this fuel (see Table 1), where it can be seen that methane is the main component of this fuel. In order to take into account the natural gas composition, species $N_2$, $CO_2$, $CH_4$ and $C_2H_6$ were introduced as input with the percentages shown in Table 1. Furthermore, the mole fraction of these species were normalized in order to have the total mole fraction equals 1.

### 3.3 $CH^*$ and $C_2^*$ reactions

As $CH^*$ and $C_2^*$ species are not present in Lindstedt et al. [10] mechanism, chemical reactions describing the production, quenching and chemiluminescence of $CH^*$ and $C_2^*$ were added to this mechanism. Also species $C_3$ was introduced since it appears in $C_2^*$ production reactions.

As expected, due to the low concentration of the excited species, no significant difference in the results was found with the addition of these elementary reactions. The thermochemical data for $CH^*$ and $C_2^*$ were added as well. The transport coefficients for the excited species were the same as those of the ground state species.

For species $CH^*$ and specially for $C_2^*$ literature data is very scarce. However, Kathrotia et al. [11] recently presented production and destruction reactions for these species. Therefore, [11] was chosen for estimating $CH^*$ and $C_2^*$ species concentration.

### 4 RESULTS

#### 4.1 Natural gas flame structure

Figure 5 presents a typical natural gas flame structure. The boundary conditions correspond to the flame I (see Section 2.2). In this figure $z = -10$ corresponds to the upper burner edge and $z = 10$ to the lower burner edge. The diffusion natural gas/air flame front is located at about $z = 0.1 \, mm$. At this point, the gas temperature passes through a maximum (1820 K).

In Figure 5 (left), methane ($CH_4$) species profile remain constant until $z = 3 \, mm$, then decreases rapidly before reaching the natural gas/air flame front. On the left side of the Figure, Oxygen ($O_2$) species profile remain constant until $z = -2 \, mm$ and then also decreases rapidly before reaching the natural gas/air flame front. In this type of flame, molecular diffusion between species is the most important phenomenon.
In Figure 5 (right), water vapor ($H_2O$) species profile is presented, and as it can be seen the profile is very similar to that of temperature. $H_2O$ is one of the main products of natural gas combustion. In hydrocarbons flames reactions usually begin with hydrogen (H) abstraction. In natural gas, methane ($CH_4$) is the main component, therefore hydrogen (H) abstraction produces methyl radical ($CH_3$). Consequently, as $CH_4$ starts to decrease at about $z = 3\ mm$ (Figure 5 (left)), $CH_3$ starts to increase. Methyl radical in turn decomposes into other radicals, explaining its presence in a thin region.

4.2 Experimental and numerical $CH^*$ and $C_2^*$ axial profiles

Figure 6 gives a comparison between experimental and numerical profiles of $CH^*$ (left) and $C_2^*$ (right) radicals for the natural gas/air Flame I. As the measurements do not give absolute values, in order to make comparisons with the numerical results, the experimental profiles of $CH^*$ and $C_2^*$ were normalized by their maximum values.

In Figure 7 is presented the same comparison but with the profiles of Flame II.

As we can see in Figures 6 and 7 (left), the numerical predictions of $CH^*$ are very similar to the experimental profiles along the axis. The positions of the peaks match almost perfectly. The experimental profiles are slightly thicker than the simulated ones, however this can be explained by the poorer spatial resolution in the experiments.

However, as we can see in Figures 6 and 7 (right), the numerical predictions of $C_2^*$ are slightly displaced compared to those of the experimental profiles along the axis. For Flame I, the difference between the numerical and experimental $C_2^*$ is less than 1 mm. For Flame II, the difference is slightly higher than 1 mm. Considering the experimental errors presented in Sections 2.2 and 2.3 and taking into account the distance between the two
Figure 6: Experimental and numerical $CH^*$ (left) and $C_2^*$ (right) species profiles - Flame I ($\phi = 4$ and $\epsilon = 134 \, s^{-1}$)

Figure 7: Experimental and numerical $CH^*$ (left) and $C_2^*$ (right) species profiles - Flame II ($\phi = 3.5$ and $\epsilon = 136 \, s^{-1}$)
burners (20 mm), these differences are not very important, therefore this good agreement between experimental and numerical \( CH^* \) and \( C_2^* \) radicals validates the chosen kinetic mechanism, and consequently the reactions of these radicals added to it.

5 CONCLUSIONS

- Paraguayan Chaco natural gas combustion in a laminar counterflow diffusion flame configuration at atmospheric pressure was studied.
- The visible chemiluminescence of the excited radicals \( CH^* \) and \( C_2^* \) is employed experimentally.
- Numerically, a 1D simulation was carried out using a standard kinetic mechanism, to which we add \( CH^* \) and \( C_2^* \) reactions.
- Typical natural gas flame structure was presented.
- A comparison between experimental and numerical profiles of \( CH^* \) and \( C_2^* \) radicals for the natural gas/air was done for two operating conditions, validating the model chosen in this study.

REFERENCES


SHAPE OPTIMIZATION OF AN AXISYMMETRIC BLUNT BODY IN HYPERSONIC FLOW FOR REDUCING DRAG AND HEAT TRANSFER

RAMESH K. AGARWAL AND CHRISTOPHER SEAGER

Washington University in St. Louis
1 Brookings Drive, St. Louis, MO 63130, USA
Email: rka@wustl.edu, URL: mems.wustl.edu

Key words: Hypersonic Flow, Shape Optimization, Drag, Heat Transfer

Abstract. A large design concern for high-speed vehicles such as next generation launch vehicles or reusable spacecraft is the drag and heat transfer experienced at hypersonic velocities. In this paper, the optimized shapes for minimum drag and heat transfer for an axisymmetric blunt body are developed using computational fluid dynamics (CFD) software in conjunction with a genetic algorithm (GA). For flow field calculations, the commercial flow solver ANSYS FLUENT is employed to solve the unsteady compressible Reynolds Averaged Navier-Stokes (RANS) equations in conjunction with the SST k-ω turbulence model. The hypersonic body shape is optimized using a multi-objective genetic algorithm (MOGA) to minimize both the drag and heat transfer. The MOGA creates a Pareto-optimal front containing the optimized shapes for various relative objectives of minimized drag and heat transfer. The results show a significant decrease in both the drag and heat transfer and exhibit the expected changes in the body profile. It should be noted that such results on shape optimization of a blunt body in hypersonic flow for reducing both drag and heat flux are reported in this paper for the first time in the literature. The proposed methodology will allow the simulation and optimization of more complex shapes for hypersonic vehicles.

1 INTRODUCTION

Hypersonic vehicles experience large drag forces and high rates of heat transfer during flight, as is the case during atmospheric re-entry. Due to the engineering challenges associated with handling these high drag forces and heat fluxes, changing the vehicle shape to reduce drag and heat transfer is a primary concern in the design of these vehicles. The cost and technical issues associated with hypersonic wind tunnel testing make routine ground testing impractical, and the experimental data from actual flight vehicles at high Mach numbers is challenging and difficult to obtain. This can be noted by the high cost and timeline of NASA’s X series of vehicles [1]. Because of these limitations on physical testing, CFD technology provides an attractive alternative to aid in the design of future hypersonic vehicles.

In this paper, the commercial CFD solver ANSYS FLUENT is employed to solve the RANS equations in conjunction with the SST k-ω turbulence model. The implementation of a turbulence model with RANS equations, when compared to the inviscid or laminar flow models, only moderately affects the drag coefficient but is extremely important in estimating the wall heat flux [2]. Since the body shape is optimized for both minimum drag and heat...
transfer, error in estimation of either drag or heat flux can lead to error in the optimized shape. Therefore, both the use of an appropriate turbulence model and mesh independence of the computed solution are vital to the accurate flow field simulation and optimization of the body shape.

The shape optimization of a blunt body for both minimum drag and heat transfer is accomplished using a genetic algorithm. A genetic algorithm (GA) is an optimization tool based on biological evolution. It iterates on individual body shapes over several generations, discarding the worst shapes and saving and altering the best. It should be noted that the GA should be exercised very carefully in determining the optimal shape since small changes in shape can lead to large differences in drag and heat transfer values. Thus, a well-written GA code not only provides the optimal body shape, but also is efficient in decreasing the total computational time needed for convergence to the optimal shape.

2 BRIEF REVIEW OF HYPERSONIC FLOW CHARACTERISTICSGENERAL SPECIFICATIONS

Hypersonic flow is characterized by many complex physical phenomena that make the aerodynamic design of hypersonic vehicles difficult. The accurate description of these complex flow phenomena makes modeling and simulation of hypersonic flows quite a challenge. Strong shockwaves close to the body, shock/boundary layer interactions, transitional and turbulent flow, and molecular dissociation of air all affect the solution, and are difficult and computationally expensive to simulate. The primary focus of this paper is on computation of drag and heat transfer using the unsteady compressible RANS equations in conjunction with SST k-ω turbulence model.

Aerodynamic drag can be divided into two components - pressure drag and friction drag [3]. In supersonic flow, wave drag, or the increased pressure behind a shockwave, dominates the pressure drag. Wave drag occurs because the front of the vehicle experiences large pressures due to shockwaves, and the aft of the vehicle experiences smaller pressures due to expansion waves. Summing these pressures over the area of the body gives a net force opposite to the direction of motion. In addition, pressure drag is influenced by the shock/boundary layer interaction and flow separation. Friction drag occurs primarily due to the velocity gradient in the boundary layer next to the wall. A turbulent boundary layer has a larger velocity gradient than a laminar boundary layer, and therefore imparts a higher skin friction drag on the body. Combining wave drag with pressure drag and skin friction, the drag on a hypersonic vehicle with a turbulent boundary layer is quite large. The total drag of the vehicle can be expressed as:

\[ D = \frac{1}{2} C_D \rho \infty V^2 \infty S \]  

(1)

where \( C_D \) is the total drag coefficient, \( \rho \infty \) is the free-stream density, \( V \infty \) is the free-stream velocity, and \( S \) is a reference area (In this paper, it is the cross sectional area of the body facing the free-stream). Therefore, for a given cross sectional shape traveling at a given altitude and speed, \( C_D \) is the only parameter that can be changed to reduce drag. Reducing \( C_D \) can be accomplished by changing the shape of the body.

Heat transfer at hypersonic speeds occurs due to aerodynamic heating. The free stream contains large amount of kinetic energy that is converted into thermal energy when the
molecules pass through the shockwave. Air temperatures at the stagnation point for hypersonic vehicles can be of the order of thousands of degrees. These high temperatures, combined with a turbulent boundary layer and decreased shock standoff distance can cause large amount of heat transfer to the body [4, 5]. Re-entry vehicles have a blunt leading edge to detach the shockwave from the body and increase the shock stand-off distance, in addition to a heat resistant thermal protection system (TPS). This poses a challenge for the design of hypersonic vehicles for minimizing both drag and heat transfer, since the blunt leading edge lowers the maximum heat transfer but results in higher drag. Thus, there is a trade-off between finding the shape for minimum drag while keeping the heat transfer below a critical value.

3 BRIEF DESCRIPTION OF GENETIC ALGORITHM AND SHAPE PARAMETERIZATION

Genetic algorithms are a class of stochastic optimization algorithms inspired by biological evolution [6]. In GA, a set of input vectors, called individuals, is iterated over, successively combining traits of the best individuals until convergence is achieved [7]. In general, a GA employs the following steps:

1. Initialization: Randomly generate N individuals.
2. Evaluation: Evaluate the fitness of each individual.
3. Natural Selection: Remove a subset of the individuals (often individuals with the lowest fitness).
   a. Parent Selection: Select pairs of individuals to produce an offspring.
   b. Crossover: Exchange information between the parent individuals.
   c. Mutation: Randomly alter some small percentage of the population.
5. Check for Convergence: If the solution has converged, return the fittest individual. If the solution has not yet converged, repeat the cycle starting at step 2 with the new generation.

Multi-objective genetic algorithms (MOGAs) differ from single-objective genetic algorithm (SOGA) in that the fitness of individuals is not determined by a single objective or fitness function, but by two or more optimization objectives [8, 9]. These objectives are sometimes in conflict [10]. This is the case in this paper, since the objective of minimizing drag and heat flux is difficult to achieve simultaneously. Thus, a unique solution to shape optimization depends on the relative weights assigned to each optimization objective. A MOGA implementation gives a set of optimal solutions known as a Pareto-optimal front. A Pareto-optimal solution contains the best possible solution for assigned weights to the two objectives. Each solution is optimal for the assigned weights to the two objectives, but is not necessarily the optimal solution for achieving both objectives under all weighted circumstances.

4 THE SHAPE OPTIMIZATION PROCESS

This section is split into four parts; each part explains an individual process that is implemented in accomplishing shape optimization. Figure 1 shows the general schematic of
the optimization process. The four processes are: (1) Shape generation, (2) Meshing the Body shape, (3) Performing simulation, and (4) Optimization using MOGA.

4.1 Shape Generation

Bezier curve polynomials are used to create the body profile. A Bezier curve is a parametric curve that is defined by control points, whereas the curve itself generates a smooth path between the control points. They have favorable characteristics for use in a GA since the initial and final points and slopes can be fixed, control point information between individuals is easily swapped, and changing a control point results in a similar change to the curve itself [11]. Figure 2 shows a set, or generation, of three random curves. The green dots represent the control points, whereas the red line represents the resulting curve. Note that in all three cases, the first two and last two control points are fixed to ensure a blunt leading edge and appropriate size constraints on the shape. The shape is axisymmetric about the bottom axis and traveling to the left at hypersonic speeds.

![Figure 2: Three randomly generated curves and their Bezier control points.](image)

4.2 Meshing the Body Shape

While the control points (green dots in Fig. 2) are saved for future use of the shape, the curve itself (red line in Fig. 2) is parameterized and loaded into ICEM Meshing software to create a mesh around the body shape. ICEM is used with a pre-written meshing script that creates the same mesh profile for different curve shapes; the far field boundaries remain the
same as does the boundary spacing, but the wall profile adjusts to the newly called curve. The mesh around each of the three randomly generated curves is shown in Fig. 3. They are axisymmetric about the horizontal axis, and the boundaries lie just far enough away from the bodies to capture shockwaves. The mesh is generated with high cell density near the wall for accurate calculation of drag and heat flux, and increases in coarseness beyond the expected shock location to increase the computational efficiency without compromising accuracy. The meshes shown in Fig. 3 provide mesh independent solutions while remaining computationally efficient.

4.3 CFD Simulations

The mesh generated in section 4.2 is loaded into commercial CFD software ANSYS FLUENT. FLUENT solves for the flow field and finds surface quantities such as skin friction, wall pressure coefficient, and total surface heat flux by solving the unsteady compressible RANS equations in conjunction with the SST k-ω turbulence model. The flow conditions and boundary conditions for the simulations are as given in Reference [12]. Free-stream Mach number is 7.1. Free-stream temperature is 293 K and wall temperature is kept constant at 1000 K. Far-field gauge pressure is 663 Pa and the pressure at the downstream boundary of the computational domain is the operating pressure in Pa extrapolated from the interior of the domain. The simulation process is also automated by using a script file. The script file loads the mesh, runs the simulations, and outputs the desired data, namely the drag coefficient and maximum wall heat flux. Figure 4 shows the Mach number contours around the three meshes in Fig. 3.
4.4 Implementation of MOGA

The above three steps are repeated until information for all individuals in a generation is recorded. Then the drag and heat transfer values are loaded into the MOGA, which runs evaluation, natural selection, and reproduction functions on the data. Throughout the optimization process, the objective function information for each body shape is always retained. The various steps of MOGA are described below:

(a) **Evaluation and Natural Selection**

The objective values for each curve are analyzed and compared to each other. In our case, minimization of the objective values is desired. Each objective value is sorted from best to worst, where the lowest value is the best and highest value is the worst. Plotting the two objective values on separate axis forms the Pareto front. Changing the weight of one objective over the other (e.g., dividing the fitness associated with the heat flux in half while keeping the drag the same) allows for different solutions to be obtained, that when combined create the Pareto-optimal front. Removal of the least fit half of individuals from use is called natural selection. Thus, only the fittest half of the data “survives” and has a chance to reproduce and influence the next generation in MOGA.

Looking at the randomly generated curves in Fig. 4, “Curve 1” has poor drag and heat transfer characteristics, and would likely get discarded during natural selection. If drag were the more important design objective, “Curve 2” would likely have a better fitness. On the other hand, if heat transfer was the more important design objective, “Curve 3” would likely have a better fitness. If both design objectives - drag and heat flux are weighted the same or very close to one another, then depending on the values of drag and heat transfer, either “Curve 2” or “Curve 3” could have a better fitness.

(b) **Reproduction and Mutation**

Reproduction involves parent selection, crossover, and mutation. From the remaining data, pairs of parent shapes are selected using a skewed random draw; the more fit a function is, the higher its probability of being selected. The paired curves then undergo crossover, where they exchange Bezier point information. The exchange can be a complete swap of certain points or a weighted average of the characteristics of certain points. Both the type of crossover and the type of exchange are selected at random. This process is repeated until the desired number of daughter shapes is created and a new generation is formed. The final step is mutation. Randomly selected Bezier points of randomly selected curves are randomly altered, adding more diversity into the generation to increase the range of body shapes.

(c) **Solution Convergence**

Finally, the MOGA checks if the curves obtained in a given generation satisfy the convergence criteria. The convergence criteria are based on whether the body shapes in a generation are close to each other and the difference in the objective values in that generation is within a specified tolerance. If the change in the body shape and the objective values for a few subsequent generations (usually 2 to 3) are in acceptable range, the solution is considered converged. Once the solution converges, MOGA gives the fittest (almost optimal) body shape meeting the desired objectives. If the solution has not yet converged, the new generation is loaded into the shape generating function in Fig. 1 and the process is repeated. Figure 5 shows a detailed schematic of information flow throughout the optimization process.
5. RESULTS AND DISCUSSION

5.1 Mesh Validation Study

The accuracy of any CFD simulation is largely dependent on the mesh that is used. In this shape optimization study for hypersonic flow, very small wall spacing is required close to the wall to accurately model the heat transfer; however the mesh spacing can be increased outside the boundary layer, and especially outside the region of the shock [13]. An extensive mesh validation study was conducted by creating an extremely dense uniform mesh around both a spherical and a randomly generated shape. The simulations were run using a 3rd order accurate scheme in FLUENT; the nearly “exact” solution was obtained for the case of sphere using a 631x1250 size mesh as shown in Table 2.

These fine meshes are computationally very expensive and the GA requires simulating at least 20 individuals over several dozens of generations. Therefore it is not practical to use these dense meshes given the available computation power. Thus, this mesh validation study focused on finding a suitable relatively coarse mesh which could give comparable results that were obtained using the fine mesh. On such a mesh, it turned out that the resulting objective values obtained from MOGA differed only slightly from the highly dense computations. Table 2 provides details of the variety of meshes used and their resulting objective values (only the values for the simulations past a spherical shape are shown). Mesh types include uniform, exponential, and segmented construction along with varied wall and far field refinements. The study determined the mesh that gave the lowest error in drag and heat transfer values, but with a runtime suitable for optimization (the 35-105-10x400 segmented mesh in Table 2). The details of this mesh are given in Figure 5.
Table 2. Mesh Validation Study

<table>
<thead>
<tr>
<th>Type of Mesh</th>
<th>Spherical Geometry</th>
<th>Mesh Size</th>
<th>Drag Coefficient</th>
<th>CD error</th>
<th>Max Heat Transfer (W/m²)</th>
<th>Max HT error</th>
<th>Max y+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td>80x200</td>
<td>.95574</td>
<td>.1</td>
<td>3080000</td>
<td>29.3</td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>120x200</td>
<td>.95534</td>
<td>.1</td>
<td>2820000</td>
<td>18.6</td>
<td>2.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>160x400</td>
<td>.95511</td>
<td>.1</td>
<td>2600000</td>
<td>9.3</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>200x400</td>
<td>.95503</td>
<td>0</td>
<td>2590000</td>
<td>8.8</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>631x1250</td>
<td>.95463</td>
<td>-----------</td>
<td>2380000</td>
<td>-----------</td>
<td>.53</td>
<td></td>
</tr>
<tr>
<td>Exponential</td>
<td>100x200 (1e-5m wall distance)</td>
<td>.95503</td>
<td>0</td>
<td>2850000</td>
<td>20</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100x200 (5e-6m wall distance)</td>
<td>.95495</td>
<td>0</td>
<td>2900000</td>
<td>22</td>
<td>.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100x200 (1e-6m wall distance)</td>
<td>.95545</td>
<td>.1</td>
<td>2900000</td>
<td>22</td>
<td>.11</td>
<td></td>
</tr>
<tr>
<td>Segmented</td>
<td>390-10x400</td>
<td>.95455</td>
<td>0</td>
<td>2470000</td>
<td>4</td>
<td>.57</td>
<td></td>
</tr>
<tr>
<td></td>
<td>85-115x400</td>
<td>.9547</td>
<td>0</td>
<td>2580000</td>
<td>8</td>
<td>.52</td>
<td></td>
</tr>
<tr>
<td></td>
<td>44-146-10x400</td>
<td>.95503</td>
<td>0</td>
<td>2550000</td>
<td>7</td>
<td>.52</td>
<td></td>
</tr>
<tr>
<td></td>
<td>25-75-10x400</td>
<td>.95464</td>
<td>0</td>
<td>2580000</td>
<td>8</td>
<td>.57</td>
<td></td>
</tr>
<tr>
<td></td>
<td>35-105-10x400</td>
<td>.95471</td>
<td>0</td>
<td>2560000</td>
<td>7</td>
<td>.41</td>
<td></td>
</tr>
<tr>
<td></td>
<td>35-105-10x250</td>
<td>.95487</td>
<td>0</td>
<td>2690000</td>
<td>13</td>
<td>.4</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5: Segmented mesh used in computation of case 35-105-10x400 in Table 2.
5.2 Shape Optimization Results with Different Weights (importance) Assigned to Drag and Heat-Flux

In Figs. 6, 7 and 8, shape optimization of the spherical nose is conducted by assigning the relative weights to drag and maximum surface heat-flux on the body in the ratio (1:1), (0.5:1) and (0.75, 1) respectively. Figure 6 shows the results obtained for equally weighted fitness functions (1:1) for drag and heat flux for flow past an optimized body.

![Figure 6](image)

**Figure 6:** Results for an optimized body with equal weighting of drag and heat transfer: Static pressure contours, wall skin friction, pressure coefficient, and wall heat flux, and convergence graphs.

Drag Coefficient: 0.745 (22% reduction relative to spherical nose)
Maximum Heat Transfer: 2230000 W/m² (13% reduction relative to spherical nose)
Figure 7 shows the results obtained for weighted fitness functions in the ratio (0.5: 1) for drag and heat flux for flow past an optimized body.

**Figure 7:** Results for an optimized body with (0.5 :1) weighting of drag and heat transfer: Mach number contours, wall skin friction, pressure coefficient, and wall heat flux.

- Drag Coefficient: 0.960 (1% increase compared to spherical nose)
- Maximum Heat Transfer: 1900000 W/m² (26% reduction compared to spherical nose)
Figure 8 shows the results obtained for weighted fitness functions in the ratio (0.75:1) for drag and heat flux for flow past an optimized body.

**Figure 8:** Results for an optimized body with (0.75:1) weighting of drag and heat transfer: Mach number contours, wall skin friction, pressure coefficient, and wall heat flux.

- Drag Coefficient: 0.825 (13% reduction compared to spherical nose)
- Maximum Heat Transfer: 2070000 W/m² (19% reduction compared to spherical nose)

### 5.3 Creation of a Pareto-Optimal Front

Combining the results from part “B” above, a Pareto-optimal front can be created as shown in Fig. 9.

**Figure 9:** Pareto-optimal front with design parameter comparisons and shape comparisons.
5.4 Discussion

The results from each individual simulation in GA were reasonable and plausible. The shockwaves were well within the outer-boundary of the mesh. Thus, the computational domain was sufficient so as not to influence the accuracy of the simulations. A great deal of effort was spent in creating a suitable mesh which was efficient but gave results which could be considered as mesh independent. It was very important for the accurate calculation of the heat flux.

The shape optimization results were obtained by assigning different weights to the importance of reducing drag vs. heat flux. The case with 100% weight assigned to drag matched well with the results from Reference [12] and the case with 50% weight assigned to drag and 50% weight assigned to heat flux matched well with the results in Reference [14]. Reference [15] shows experimentally that a blunted cone shape has lower drag than a standard cone at supersonic speeds, and the optimized shapes appear to be approaching a blunted cone shape with a bluntness ratio of around 0.4 as the importance of reducing the drag is increased. The optimized shape found in Reference [12] focused only on drag and not on heat transfer. In our simulations, as the importance of reducing drag increases relative to maximum heat flux, our optimized shapes appear to approach similar to that in Reference [12]. Heat transfer values become higher, but are of similar magnitude to those reported in Reference [14]. This is because the turbulent flow leads to greater heat transfer than laminar flow.

The multi-objective optimization results for reducing both the drag and heat flux are obtained for the first time in the literature. Heat transfer decreases more than the drag coefficient as the weighting is increased for the heat flux. The change in geometry occurs as expected with the optimized shapes having increasingly larger leading edge radius of curvature since the heat transfer begins to dominate the optimization criteria. Nevertheless, the curves maintain a streamlined shape to keep the drag from becoming too large. Additionally, the plot of the Pareto-optimal front shows a well-fitted curve connecting the three cases considered using different weights for drag and heat flux; it can be extrapolated backwards for drag-favored optimization as shown in Fig. 9. Previous simulations support the approximate value of the drag obtained from extrapolation, since previous drag-only optimizations converged to a drag coefficient value of about 0.63. Further simulations are currently being conducted on finding the shapes of drag-favored optimization using the MOGA with the expectation that they would lie close to the Pareto-optimal front.

6. CONCLUSIONS AND FUTURE WORK

This study focused on using a multi-objective genetic algorithm (MOGA) to minimize both the drag coefficient and the maximum heat flux for an axisymmetric blunt body in hypersonic flow. Bezier polynomials were used to create shape profiles that were meshed and solved by the commercial CFD solver ANSYS FLUENT. FLUENT solved the flow field of each shape by solving the compressible Reynolds-averaged Navier-Stokes equations in conjunction with the SST k-ω turbulence model. A comprehensive mesh validation study was conducted to determine an efficient segmented mesh that could be used to obtain mesh independent solutions. Individual results of simulations for three cases employing different weights to drag and heat flux as fitness functions were analyzed before being combined into a Pareto-optimal front. The Pareto optimal front displayed expected characteristics, e.g. the...
leading edge radius of curvature of the optimized shape increased if higher weight was assigned to heat flux and vice versa. This paper shows that a MOGA can be employed as a practical tool for optimizing blunt body shapes in hypersonic flow for reducing both the drag and heat transfer. This research shows that to achieve two objectives, several options are available to the designer depending upon the relative importance of the desired objective. In case of drag-favored optimization (reducing drag being the more important objective), higher weight could be assigned to minimization of drag compared to minimization of heat transfer. Such results on shape optimization of a blunt body in hypersonic flow for reducing both the drag and heat transfer have never been reported in the literature before. Future work may include optimizing the current blunt body using dissociated and rarefied air models, using different transitional/turbulence models depending upon the altitude and optimization of other axisymmetric shapes such as a bi-cone and three-dimensional non-axisymmetric shapes.

REFERENCES

A NOVEL ALGEBRAIC EXTRA-STRESS MODEL AND ITS ANALYTIC SOLUTION FOR FULLY DEVELOPED LAMINAR FLOWS

GILCILENE S. PAULO, GILMAR MOMPEAN, CASSIO M. OISHI AND MURILO F. TOMÉ

1Faculdade de Ciências e Tecnologia (FCT), UNESP - Univ Estadual Paulista
Rua Roberto Simonsen 305, 19060-900 Presidente Prudente/SP, Brazil
gilcilene@fct.unesp.br, oishi@fct.unesp.br

2Université des Sciences et Technologies de Lille, Laboratoire de Mécanique de Lille,
URM-CNRS 8107
Cité Scientifique, 59655 Villeneuve d’Ascq Cedex, France
gilmar.mompean@polytech-lille.fr

3Departamento de Matemática Aplicada e Estatística, Universidade de São Paulo, USP
Av. Trabalhador So-carlense, 400, 13566-590 São Carlos/SP, Brazil
murilo@icmc.usp.br

Key words: FENE-CR Model, Algebraic Model, Analytic Solution, Finite Difference Method

Abstract. This work presents the development of an algebraic extra-stress model (AESM) for modelling viscoelastic flows governed by the constitutive Finite Extendable Non-linear Elastic - Chilcott and Rallison (FENE-CR) differential model. An algorithm for solving the governing equations for incompressible flows of FENE-CR fluids by the AESM model is given. Moreover, by considering fully developed flows, analytic expressions for the conformation tensor A and velocity are derived. These analytic solutions are employed to verify the convergence and accuracy of the numerical methodology for simulating flows modelled by this new AESM.

1 INTRODUCTION

Two-dimensional flows described by differential models employ three partial differential equations for calculating the components of the extra-stress tensor whereas an algebraic model requires only one partial differential equation to be solved. Therefore, algebraic models are attractive because their solutions demand less CPU time and the algebraic equations makes the computational algorithms easier to implement.
The first AESM was developed by Mompean and Deville [1] who used the ideas for modeling turbulent flows of Newtonian fluids and derived an AESM model to obtain approximate solutions of flows described by the Oldroyd-B differential model. In that work, numerical results were performed in order to demonstrate the abilities of the formulation. In a subsequent paper, Mompean [2] extended the idea the Oldroyd-B AESM to approximate the differential PTT model. In this work, we use the ideas presented in the methodology proposed in [1, 2], and derive an AESM to approximate the FENE-CR differential model.

The system of dimensionless equations representing isothermal, incompressible, viscoelastic flows governed by the FENE-CR constitutive equation can be written as

\[
\frac{\partial \mathbf{u}}{\partial t} + \nabla \cdot (\mathbf{u} \mathbf{u}) = -\nabla p + \frac{\beta}{Re} \nabla^2 \mathbf{u} + \nabla \cdot \tau,
\]

\[
\nabla \cdot \mathbf{u} = 0,
\]

\[
\frac{DA}{Dt} - \nabla \mathbf{u} \cdot \mathbf{A} - \mathbf{A} \cdot (\nabla \mathbf{u})^T = -\frac{1}{Wi} f(I_A) [\mathbf{A} - I],
\]

\[
\tau = \frac{(1 - \beta)}{ReWi} f(I_A) [\mathbf{A} - I].
\]

In these equations, the function \(f(I_A) = L^2/(L^2 - I_A)\) is positive, where \(L^2 > 3\) is the finite extensibility of the polymer molecules parameter, \(t\) is the time, \(\mathbf{u}\) is the velocity field, \(p\) is the pressure, \(\tau\) and \(I_A\) are the extra-stress tensor and the first invariant (the trace) of the conformation tensor \(\mathbf{A}\), respectively. The parameter \(\beta\) controls the amount of Newtonian solvent contribution, \(\beta = \eta_S/\eta_0\), where \(\eta_0\) is the zero shear viscosity. The nondimensional numbers \(Re = \rho U H/\eta_0\) and \(Wi = \lambda U / H\) are, respectively, the Reynolds and the Weissenberg numbers, where \(H\) and \(U\) denote typical length and velocity scalings. The symbol \(D/Dt\) represents the material derivative.

For further information about notation, in this paper either \(I_\Upsilon\) or \(\{\Upsilon\}\) can be used for the trace of a generic tensor \(\Upsilon\).

2 FENE-CR ALGEBRAIC EXTRA-STRESS MODEL

By following the ideas of Mompean [1, 2], we use the vorticity tensor \(\mathbf{W} = \frac{1}{2} (\nabla \mathbf{u} - (\nabla \mathbf{u})^\top)\) and the rate of deformation tensor \(\mathbf{S} = \frac{1}{2} (\nabla \mathbf{u} + (\nabla \mathbf{u})^\top)\) and rewrite Eq. (3) as

\[
\frac{DA}{Dt} = -\frac{f(I_A)}{Wi} (\mathbf{A} - I) + (\mathbf{S} \mathbf{A} + \mathbf{A} \mathbf{S}) - (\mathbf{A} \mathbf{W} - \mathbf{W} \mathbf{A}).
\]

For the development of an AESM, it is necessary to work with the deviatoric tensor,

\[
\Gamma_A = \mathbf{A} - \frac{I_A}{3} I.
\]
An evolution equation for $I_A$ is obtained by taking the trace in Eq. (5) and after a straightforward calculation, we obtain

$$\frac{DI_A}{Dt} = -\frac{f(I_A)}{Wi}(I_A - 3) + 2\{\Gamma_A S\}$$

(7)

and using Eqs. (6) and (7), equation (5) can be written in terms of $\Gamma_A$ as

$$\frac{D\Gamma_A}{Dt} = -\frac{f(I_A)}{Wi} \Gamma_A + \left( S\Gamma_A + \Gamma_A S - \frac{2}{3}\{\Gamma_A S\} I \right) - (\Gamma_A W - W\Gamma_A) + \frac{2}{3} I_A S.$$  

(8)

### 2.1 Assumption

In two-dimensional flows, the tensor $\Gamma_A$ can be represented as,

$$\Gamma_A = \frac{\{\Gamma_A S\}}{\{S^2\}} S + \frac{\{\Gamma_A W S\}}{\{S^2\}\{W^2\}} (SW - WS) + \frac{6\{\Gamma_A S^2\}}{\{S^2\}^2} \left( S^2 - \frac{1}{3}\{S^2\} I \right),$$

(9)

since this tensor is symmetric and traceless. For details, see [3].

Following the work of Mompean [1, 2], the idea is to use Eq. (8) to obtain an explicit expression for $\Gamma_A$. For this, Eq. (8) is approximated by

$$\frac{D\Gamma_A}{Dt} = \alpha \Gamma_A,$$

(10)

where for the FENE-CR model, it is found that $\alpha = \frac{1}{(I_A - 3)} \frac{DI_A}{Dt}$ that makes $\tau_{zz} = 0$, which is required in two-dimensional flows.

By introducing Eq. (10) into Eq. (8) and using Eq. (7) for $DI_A/Dt$, yields

$$0 = -\frac{2}{(I_A - 3)}\{\Gamma_A S\} \Gamma_A + \left( S\Gamma_A + \Gamma_A S - \frac{2}{3}\{\Gamma_A S\} I \right) - (\Gamma_A W - W\Gamma_A) + \frac{2}{3} I_A S.$$  

(11)

From Eq. (11) we obtain expressions for $\{\Gamma_A S\}$, $\{\Gamma_A S^2\}$ and $\{\Gamma_A W S\}$. For instance, the term $\{\Gamma_A S\}$ is obtained by multiplying Eq. (11) by $S$ and take the trace in the resulting expression. The other terms are constructed in a similar manner. Thus, the expressions obtained for the scalar coefficients are

$$\{\Gamma_A S\} = \sqrt{0.5(I_A - 3) \left[ (I_A - 1)\{S^2\} + (I_A - 3)\{W^2\} \right]},$$

$$\{\Gamma_A S^2\} = \frac{(I_A - 3)}{6}\{S^2\},$$

$$\{\Gamma_A W S\} = -\frac{(I_A - 3)}{2}\{W^2\}.$$  

(12)

Finally, substituting the expressions in Eq. (12) into Eq. (9), we obtain the explicit expression of $\Gamma_A$ for the FENE-CR model:

$$\Gamma_A = \sqrt{0.5(I_A - 3) \left[ (I_A - 1)\{S^2\} + (I_A - 3)\{W^2\} \right]} S$$

$$- \frac{1}{2}\frac{(I_A - 3)}{\{S^2\}} (SW - WS) + \frac{(I_A - 3)}{\{S^2\}} \left( S^2 - \frac{1}{3}\{S^2\} I \right).$$

(13)
In summary, the numerical solution of the FENE-CR AESM can be obtained by the following steps: (i) use Eq. (12) to calculate \(\{\Gamma_A S\}\), (ii) solve Eq. (7) for \(I_A\), (iii) calculate \(\Gamma_A\) from Eq. (13), (iv) apply Eq. (6) to recover the conformation tensor \(A\), and finally, (v) update \(\tau\) from Eq. (4).

3 ANALYTIC SOLUTION OF THE FENE-CR AESM FOR FULLY DEVELOPED CHANNEL FLOWS

Consider the flow in a two-dimensional channel of width \(-0.5 \leq y \leq 0.5\). The assumptions of fully developed flow lead to the following

\[
\begin{align*}
  u &= u(y), \quad v = 0, \quad I_A = I_A(y), \quad \frac{\partial (\cdot)}{\partial t} = \frac{\partial (\cdot)}{\partial x} = 0, \quad \frac{\partial p}{\partial x} = p_x \text{ (constant)}; \\
  \frac{\partial p}{\partial y} &= 0.
\end{align*}
\]

which after been introduced into Eqs. (1), (7) and (13), produces

\[
\begin{align*}
  0 &= -p_x + \frac{\beta}{Re} \frac{d^2 u}{dy^2} + \frac{\partial \tau_{xy}}{\partial y}, \\
  0 &= -\frac{f(I_A)}{Wi} (I_A - 3) + 2\sqrt{0.5(I_A - 3)(du/dy)^2}, \\
  \Gamma_A^{xx} &= 2/3 (I_A - 3), \quad \Gamma_A^{yy} = -1/3 (I_A - 3), \quad \Gamma_A^{xy} = \sqrt{0.5(I_A - 3)}.
\end{align*}
\]

Solving Eq. (16) for \(I_A\), the following quadratic equation is obtained

\[
\begin{align*}
  a I_A^2 + b I_A + c &= 0,
\end{align*}
\]

where \(a = 2Wi^2 (du/dy)^2\), \(b = -L^4 - 4L^2Wi^2 (du/dy)^2\) and \(c = 3L^4 + 2L^4Wi^2 (du/dy)^2\).

Since \(L^2 > 3\), its discriminant \(\Delta = L^8 + 8L^4Wi^2 (du/dy)^2 (L^2 - 3)\) is positive, hence, there are two different solutions for the variable \(I_A\). An important theoretical question is the correct choice of the roots obtained in the solution of the Eq. (18). To decide which root should be disregarded, we proceed as follows.

Notice that if we consider the expression \(I_A = -\frac{b}{2a} + \frac{\sqrt{\Delta}}{2a}\), then we obtain \(I_A = L^2 + \frac{L^4 + \sqrt{\Delta}}{4Wi^2 (du/dy)^2} > L^2\), which is an absurd since the definition of \(f(I_A)\) assumes \(I_A < L^2\), whereas the other expression

\[
I_A = -\frac{b}{2a} - \frac{\sqrt{\Delta}}{2a}
\]

satisfies the correct assumption of the definition of \(f\). Indeed, the solution given by Eq. (19) for fully-developed channel flow of the FENE-CR AESM coincides with the trace of the conformation tensor \(A\), for the same flow, whether the FENE-CR differential model is considered (see Appendix A from [4]). Moreover, by taking the Eqs. (6) and (17) into account, we also obtain the same solution as obtained in [4] for the components \(A^{xx}\), \(A^{yy}\) and \(A^{xy}\). Consequently, from Eq. (4), the components of the non-Newtonian extra-stress tensor, \(\tau^{xx}\), \(\tau^{yy}\) and \(\tau^{xy}\), are the same for differential and algebraic FENE-CR models.
4 NUMERICAL METHODOLOGY

The numerical method is based on the GENSMAC methodology for simulating viscoelastic flows developed by Tomé and co-workers. [5, 6].

The basic equations are approximated on a uniform 2D Cartesian staggered grid using the finite difference method. To calculate the numerical solution at time $t_{n+1} = t_n + \delta t$, the momentum equation is solved by an implicit method in which the velocity and pressure are treated implicitly. This calculation follows the ideas behind the projection method (see in [7]) that decouples velocity and pressure in the Navier-Stokes equations. The temporal evolution for $I_A$ is obtained by an explicit 2nd-order Runge-Kutta method. The high order CUBISTA method [8] is employed for solving the advection terms while the spatial derivatives are approximated by 2nd. order finite differences.

The numerical solutions are obtained from the initial conditions $u = 0$ and $I_A = 3$ and the following boundary conditions: on the rigid boundary is imposed $I_A = 3$ at the first computational cycle, thereafter an extrapolation that is second-order accurate is applied; for the velocity field the no-slip condition is imposed $u = 0$. The homogeneous Neumann condition is employed at the exit boundary for both $u$ and $I_A$. At the inlet boundary is either imposed the fully-developed flow profile, as described in section 3, or a constant profile that is related to the problem that is being solved. Thus, the following steps are performed at each computational cycle:

**Step 1:** Solve Eq. (7) to obtain $I_A^{(n+1)}$:

$$\frac{I_A^{(n+1)} - I_A^n}{dt} = -\nabla \cdot (u I_A^n) - f(I_A^n)(I_A^n - 3) + 2\{\Gamma_A S\}^n,$$

where $\{\Gamma_A S\}^n = 0.5(I_A^n - 3) \cdot ([(I_A^n - 1)\{S^2\}^n + (I_A^n - 3)\{W^2\}^n] .

**Step 2:** Calculate $\Gamma_A^{(n+1)}$ from the algebraic equation (13):

$$\Gamma_A^{(n+1)} = \{\Gamma_A S\}^{(n+1)} = \{S^2\}^n - \frac{1}{2}(I_A^{(n+1)} - 3)\{(S^2 - W S)^2\}^n + \frac{I_A^{(n+1)} - 3}{3\{S^2\}^n} \left( S^2 - \frac{1}{3}\{S^2\}\hat{I} \right)^n,$$

where $\{\Gamma_A S\}^{(n+1)} = \sqrt{0.5(I_A^{(n+1)} - 3) \cdot [(I_A^{(n+1)} - 1)\{S^2\}^n + (I_A^{(n+1)} - 3)\{W^2\}^n] .

**Step 3:** Calculate the components of the conformation tensor $A^{(n+1)}$ from Eq. (6):

$$A^{(n+1)} = \Gamma_A^{(n+1)} + \frac{1}{3} I_A^{(n+1)} \hat{I}.$$
Step 4: Compute the components of the non-Newtonian tensor $\tau^{(n+1)}$ from Eq. (4):
\[
\tau^{(n+1)} = \frac{(1 - \beta)}{Re W_i} f(I^{(n+1)}_A)[A^{(n+1)} - I].
\] (23)

Step 5: Calculate a provisional velocity field $\tilde{u}$ by an implicit time discretization:
\[
\frac{\tilde{u}}{\delta t} - \frac{\beta}{Re} \nabla^2 \tilde{u} = \frac{u^n}{\delta t} + \nabla \cdot (uu)^n - \nabla p^n + \nabla \cdot \tau^{(n+1)},
\] (24)
subjected to the same boundary conditions obeyed by $u$.

Step 6: Solve the Poisson equation for the potential function $\psi$:
\[
\nabla^2 \psi^{(n+1)} = \nabla \cdot \tilde{u}.
\] (25)

The boundary conditions required for solving this Poisson equation are the homogeneous Neumann conditions for rigid walls and inlets, while homogeneous Dirichlet conditions are applied at outlets.

Step 7: Compute the final velocity field:
\[
u^{(n+1)} = \tilde{u} - \nabla \psi^{(n+1)}.
\] (26)

Step 8: Compute the final pressure field from equation deduced from the projection method [7]:
\[
p^{(n+1)} = p^n + \frac{\psi^{(n+1)}}{\delta t} - \frac{\beta}{Re} \nabla^2 \psi^{(n+1)}.
\] (27)

Step 9: Update the calculations of the trace of the conformation tensor $I_A$, and of the provisional velocity vector $\tilde{u}$ (by using $u$) on all boundaries.

4.1 Numerical verification

The analytical expressions derived in Section 3 are applied to verify the correctness and accuracy of the numerical method developed for solving the FENE-CR AESM, as follows.

The steady state flow in a horizontal channel was simulated. The initial conditions were $u = 0$ and $I_A = 3$. The boundary conditions on rigid boundaries and at the outflow were those described in Section 4. On the inlet, the analytic profiles derived in Section 3 were imposed. The following data were employed: inlet width $H = 1$m, channel length
10H, Re = 0.1, Wi = 0.3, β = 0.8, $L^2 = 200$. The channel flow was simulated on the three meshes: M20: 200×20 cells, M40: 400×40 cells and M80: 800×80 cells.

After the steady state was reached, the numerical solutions on the cross section at the middle of the channel ($x = 5$) were considered where the relative errors between the exact and the numerical values were computed by the $l_2$ norm. The results are displayed in Fig. 1, in log-log scales where it can be seen that the spatial order of the numerical methodology is approximately 2.

![Figure 1](image)

Figure 1: Variation of the relative errors with mesh refinement.

5 CONCLUSIONS

This work presented a novel algebraic extra-stress model to simulate viscoelastic flows governed by the FENE-CR model. By using this model, instead of solving separately for each component of the conformation tensor, the algebraic formulation solves only one equation for the trace of the conformation tensor. Moreover, analytic solutions for the algebraic extra-stress model in fully developed channel flow were derived. It was found that the analytic solutions for velocity and pressure were found to be the same solutions given in [4], demonstrating that this novel algebraic model is capable of simulating shear flows defined by the classic differential FENE-CR model. In addition, a numerical methodology to solve the flow of a FENE-CR AESM fluid was implemented and applied to solve the flow in a two-dimensional channel. The analytic solutions were employed to verify the correctness and accuracy of the proposed numerical method that presented a convergence
order of 2. Future works include to apply this new algebraic extra-stress model to simulate flows in complex geometries and to test its abilities to predict viscoelastic flows at high Weissenberg numbers.

REFERENCES


SOLUTION OF TWO-DIMENSIONAL SHALLOW WATER EQUATIONS BY A LOCALIZED RADIAL BASIS FUNCTION COLLOCATION METHOD

CARLOS A. BUSTAMANTE, HENRY POWER, CESAR NIETO AND WHADY F. FLOREZ

1School of Engineering, Universidad Pontificia Bolivariana  
Circular 1 No 70-34, Medellin, Colombia  
carlos.bustamante@upb.edu.co, cesar.nieto@upb.edu.co, whady.florez@upb.edu.co, www.upb.edu.co

2School of Mechanical, Materials and Manufacturing Engineering, University of Nottingham  
University Park, Nottingham NG7 2RD, UK  
henry.power@nottingham.ac.uk, www.nottingham.ac.uk

Key words: Shallow water equations, radial basis functions, non-linear PDE system, meshless methods

Abstract. A localized version of the Radial Basis Function (RBF) collocation method, is employed for solving the two-dimensional Shallow Water Equations (SWEs). The Newton-Raphson method is used to solve the resulting system of non-linear algebraic equations and a first-order implicit scheme is implemented for temporal discretisation. In order to obtain non-smooth functions which describe height and velocity at supercritical regimes, RBFs are used to approximate dependent variable products instead variable values by means of a local formulation in generalized finite difference fashion. Accurate solutions are obtained for sub- and trans-critical regimes in one-dimensional problems whose analytical solution are reported elsewhere. For verifying two-dimensional spatial and temporal discretization, an oscillatory problem with analytical solution are solved to show the capability of the method for accurately solve 2-D problems. In general, excellent results are achieved regarding the fact that no upwinding scheme are employed when solving transcritical regime problems.

1 INTRODUCTION

The Shallow Water Equations (SWEs) has been widely used to model one- and two-dimensional open-channel flow in situations whereby depth is much smaller than wide such as rivers, bays and straits. For the two-dimensional case, SWEs consists in a system of non-linear equations in terms of depth-average velocity vector and water surface height.
Additionally, in most of the practical problems the SWEs must be solved in complex domains given the geometrical irregularities presented in borders of open water systems. Therefore, the use of meshless methods for solving SWEs has been popularising during the last decades regarding the reduction in preprocessing time, the use of high-order approximations and a versatile treatment of boundary conditions. Several meshless methods has been applied to the solution of SWEs such as Radial Basis Function (RBF) direct collocation [6], Differential Quadrature (RBF-DQ) [6, 3], Natural element method (NEM) [1] and Discrete Least Square Method (DLSM) [7].

Among meshless methods, RBF direct collocation has been used to solve a variety of partial differential equation (PDE) problems such as Poisson, Helmholtz, Convection-diffusion and Navier-Stokes. Given a straightforward implementation and the excellent behaviour of RBF in interpolation, the application of direct RBF collocation for solving SWEs has been popularised recently. Hon et al. [4] successfully solved a 2-D shallow water problem in order to find heights at different points in a bay under storm conditions by using a global RBF collocation scheme. Local formulation were implemented by Sun et al. [8] and Khoshferat and Abedeni [6], the first of them validating the scheme with a two-dimensional dam-break situation and applying it to the analysis of a flood prevention system. Khoshferat and Abedeni [6] developed a hybrid formulation with RBF-DQ and local direct collocation with the aim of obtaining water height in Oresund strait. As shown above, successful simulations has been attained for solving SWEs by RBF meshless methods in situations where height is much greater than depth. Open channel and river simulations, in which supercritical conditions are common, has not been widely explored by direct RBF methods.

In this work, a simple localised RBF collocation method is implemented with the aim of solving the SWEs for one- and two-dimensional problems. Newton-Raphson method is employed to solve the resulting non-linear system of equations and an implicit first-order scheme is implemented to treat the transient term. At super-critical flow, i.e. Froude number greater than one, non-smooth behaviour of height and velocity functions are well approximated by using the RBF collocation to calculate, no height and velocity spatial derivatives, but variable product derivatives. One- and two-dimensional problems with analytical solution are solved to verify the approach and to show accuracy and applicability of the proposed method.

2 SHALLOW WATER EQUATIONS

The two-dimensional SWEs includes mass balance and the two components of the momentum conservation law, both of them in terms of the water surface height $h$ and depth-average velocity components $u$ and $v$, as follows:
\begin{align*}
\frac{\partial h}{\partial t} + \frac{\partial hu}{\partial x} + \frac{\partial hv}{\partial y} &= 0 \quad (1) \\
\frac{\partial hu}{\partial t} + \frac{\partial hu^2}{\partial x} + \frac{\partial huv}{\partial y} + gh \frac{\partial H}{\partial x} &= -ghS_{fx} \quad (2) \\
\frac{\partial hv}{\partial t} + \frac{\partial huv}{\partial x} + \frac{\partial hv^2}{\partial y} + gh \frac{\partial H}{\partial y} &= -ghS_{fy} \quad (3)
\end{align*}

Water surface height \( h \) is measured from soil bottom, whose shape is given by the topography function \( z(x, y) \), while total height is computed as \( H = h + z \). In this case, the friction force \( S_f = (S_{fx}, S_{fy}) \) is expressed according to the Manning-Striker’s law, in the following way:

\[ S_f^* = n^2 \frac{\bar{U}\sqrt{u^2 + v^2}}{h^3} \quad (4) \]

where \( \bar{U} = (u, v) \) and \( n \) is the Manning constant. Three types of boundary conditions are commonly used in SWEs: known height, known \( x \) and \( y \) discharge (\( uh \) and \( vh \), resp.) and land. In case of land boundary condition \( u = 0 \) and \( v = 0 \).

3 RBF COLLOCATION METHOD FOR SWE TRANSIENT SOLUTION

As Kansa [5] formulates, the dependent variable of a boundary value problem can be approximated as a linear combination of RBFs. When the approximation is substituted into the governing PDE and the boundary condition and the resulting expression is evaluated in a finite number of nodes scattered throughout domain and boundary, it is possible to obtain a linear system of equations whose solution are the coefficients of the original linear combination of RBFs. Regarding the fact that SWEs are a set of non-linear PDEs, RBF collocation must be implemented in conjunction to a linearization scheme or a non-linear solver. In our case, the Newton-Raphson method is implemented and an implicit scheme is applied for time discretisation.

Let us consider a generic function \( \phi \) which is approximated by RBFs, in this case the Multiquadric (MQ) function with \( m = 1 \), \( \psi = (r^2 + c^2)^{1/2} \) as follows:

\[ \phi = \sum_{i=1}^{N} \alpha_i \psi(r_i) + \sum \beta P(\tilde{x})^{m-1} \quad (5) \]

with \( N \) as the number of points \( \tilde{c} \) whereby the RBF is evaluated, regarding that \( r_i = \sqrt{(\tilde{x} - \tilde{c}_i) \cdot (\tilde{x} - \tilde{c}_i)} \). In the localised formulation, \( N \) is the number of the nearest points to the collocation point \( \tilde{x} \), i.e. nodes inside of the subdomain defined by a characteristic number of nodes or local radius. The polynomial term is added in order to ensure invertibility of resulting interpolation matrix. Given \( m = 1 \), the polynomial terms
are reduced to a single constant. Therefore, \( \phi \) spatial derivatives are calculated according to:

\[
\frac{\partial \phi}{\partial x_k} = \sum_{i=1}^{N} \alpha_i \frac{\partial \psi(r_i)}{\partial x_k}
\]  

(6)

After collocation of equation (5) at the nodes located inside a subdomain, the following linear system of equations, \([\Psi][\alpha] = [\Phi]\), is obtained,

\[
\begin{pmatrix}
\psi(r_1) & \cdots & \psi(r_N) & 1 \\
\vdots & \ddots & \vdots & \vdots \\
\psi(r_N) & \cdots & \psi(r_N) & 1 \\
1 & \cdots & 1 & 0
\end{pmatrix}
\begin{pmatrix}
\alpha_1 \\
\vdots \\
\alpha_N \\
\beta
\end{pmatrix}
= 
\begin{pmatrix}
\phi(x_1) \\
\vdots \\
\phi(x_N) \\
0
\end{pmatrix}
\]  

(7)

and unknown coefficients can be expressed in terms of unknown variable values as \([\alpha] = [\Psi]^{-1}[\Phi]\). Therefore, equation (5) is rewritten in terms of unknown variable values and matrix products, as:

\[
\phi(x) = [G(x)][\Psi]^{-1}[\Phi]
\]  

(8)

where \(1 \times N + 1\) matrix \([G(x)] = [\phi(r_1), \cdots, \phi(r_N), 1]\).

With the aim of avoiding discontinuity effects on the stability of the solution, a slightly different approximation is done here, in the sense that velocity and height derivatives are not approximated by RBFs but the spatial derivatives of variable products according to the governing equations. In order to discretise SWEs, the generic variable \(\phi_j\) takes the values reported in Table 1 in terms of the index \(j\).

<table>
<thead>
<tr>
<th>(j)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\phi_{j})</td>
<td>(hu)</td>
<td>(hv)</td>
<td>(hu^2)</td>
<td>(hv^2)</td>
<td>(huv)</td>
<td>(H)</td>
</tr>
</tbody>
</table>

If the RBF approximations are substituted into equations from (1) to (3) and the temporal derivatives are approximated by an implicit first-order scheme, a set of discretised non-linear algebraic equations, corresponding to the generic case of an interior subdomain including \(N\) nodes, is obtained. After applying those equations to each one of the internal nodes and, by replacing the corresponding one with the boundary condition, at boundary nodes, a non-linear system of \(3N_t\) equations, with \(N_t\) as the sum of internal and boundary nodes, is obtained. In the present case, the system of equation is solved by using the Newton-Raphson method for each time step.
4 NUMERICAL RESULTS

Several shallow water problems with analytical solution are solved in order to validate the implemented scheme. Accuracy of spatial discretisation is evaluated by the comparison of numerical results with the analytical solution for the 1 − D problem of flow over a bump both in sub- and super-critical regimes. Transient term discretisation is tested by analysing the error for several times of the 1 − D dam-break problem. Spatial discretisation in 2 − D situations is assessed by solving a transient problem with analytical solution.

4.1 Steady 1-D problems

Different analytical solutions of SWEs for one-dimensional problems can be obtained according to the methodology presented in [2] after defining the topography function \( z(x, y) \) and the friction term value. With the aim of testing the accuracy of the spatial discretisation, the transient code is employed to solve the steady problems starting at an initial guess. As a first validation problem, the steady and frictionless flow in a one-dimensional open channel with a bump is addressed. The domain length is \( L = 25 \) and the topography function is given by:

\[
z(x) = \begin{cases} 
0 & \text{else} \\
0.2 - 0.05(x - 10)^2 & 8 < x < 12 
\end{cases} 
\] (9)

Froude number \( Fr \), defined by equation (10), provides a relationship between inertial and gravitational forces. Additionally, when \( Fr < 1 \) the flow is sub-critical, i.e. information spreads upwind and downwind. In case \( Fr > 1 \), information only spreads downwind direction and the flow presents super-critical behaviour. Combinations of the mentioned regimes are known as transcritical flow.

\[
Fr = \frac{|\vec{U}|}{\sqrt{gh}} 
\] (10)

4.1.1 Sub-critical flow over a bump

Given the topography function (9) and subcritical flow behaviour, the following analytical solution is reported for height in [2]:

\[
h(x)^3 + \left( z(x) - \frac{p_0^2}{2ghL^2} - h_L \right) h(x)^2 + \frac{p_0^2}{2g} = 0 
\] (11)

where \( p_0 \) is a constant value for the discharge, fixed as a boundary condition, as well as the height at outflow boundary given by \( h_L = h(L) \). The initial solution corresponds to water at rest, defined by:

\[
h(x) + z(x) = 2 \quad \forall x \in [0, L] 
\] (12)

\[
h(x)u(x) = p(x) = 0 \quad \forall x \in [0, L] 
\] (13)
Constant values for boundary conditions are given by \( h(0)u(0) = p_0 = 4.42 \) and \( h(L) = 2 \). Once analytical height is obtained by solving the third grade polynomial (11), it is possible to find velocity since discharge \( p = hu \) is constant throughout domain for one-dimensional steady situations (see equation (1)). In Figure 1, analytic (Exact) and obtained numerical (Present) solutions are presented for the one-dimension sub-critical flow problem. It is solved with 81 equally-spaced nodes. Topography shows a bump with a maximum at \( x = 10 \). As can be seen in the Figure 1, the bump produces a small depression on the surface and a corresponding velocity increase. Downwind the bump, velocity and height values are the same as at the inlet boundary.

In this case, the error obtained for numerical solution (quantified as the relative \( L_2 \)-norm of error) is not comparable to the variable values and, in consequence, there are no appreciable differences in figures 1a and b.

With the aim of analysing the convergence order of the spatial discretisation scheme, the present problem is solved for different number of nodes and different values of the shape parameter \( c \) of the MQ function (Figure 2). In the first case, four nodal distribution are employed \((N = 41, 81, 161, 321)\) and, according to trend obtained, the convergence of the scheme is shown. If a potential regression is made with data presented in Figure 2, a second-order convergence rate is found (between 2.02 and 2.44). The error behaviour in terms of the shape parameter presented agrees the usual trend in direct RBF collocation methods: after a constant value zone, the error decreases when the shape parameter increases up to a critical value whereby the error increases due to the ill-conditioning of the local interpolation matrices.

Additionally, the behaviour of error and relative cpu-time spent are examined in terms of the number of nodes per sub-domain in Table 2. Relative cpu-time is computed as the time spent divided by the time needed to obtained numerical solution when \( N = 81 \) and...
Carlos A. Bustamante, Henry Power, Cesar Nieto and Whady F. Florez

\[ \Delta x_1 \]

\[ \varepsilon_{L2}^{c=0.01} \]

\[ \varepsilon_{L2}^{c=0.1} \]

\[ \varepsilon_{L2}^{c=1.0} \]

Figure 2: Relative $L_2$ norm of the height error in terms of node spacing (left) and shape parameter (right) for subcritical flow over a bump

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\varepsilon_{L2}$</th>
<th>$t_{CPU}/t_{CPU81}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>$5.2693 \times 10^{-5}$</td>
<td>0.42</td>
</tr>
<tr>
<td>4</td>
<td>$3.8908 \times 10^{-5}$</td>
<td>0.71</td>
</tr>
<tr>
<td>8</td>
<td>$1.9669 \times 10^{-5}$</td>
<td>1.00</td>
</tr>
<tr>
<td>10</td>
<td>$1.6509 \times 10^{-5}$</td>
<td>1.12</td>
</tr>
<tr>
<td>16</td>
<td>$1.1708 \times 10^{-5}$</td>
<td>1.60</td>
</tr>
</tbody>
</table>

Table 2: Relative $L_2$ norm of the height error and computational cost in terms of the number of nodes per subdomain

$n = 8$. As expected, error diminishes and cpu-time increases when the number of nodes is increased. Nevertheless, error diminishing is almost negligible for $n > 8$ while relative cpu-time preserves a similar increasing rate for all $n$ values tested. Therefore, following problems are solved with $n = 8$.

4.1.2 Transcritical flow over a bump

In this case, the analytical solution for height is modified in the following manner in order to capture behaviour at supercritical regime:

\[ h(x)^3 + \left( z(x) - \frac{p_0^2}{2gh_M^2} - h_M - z_M \right) h(x)^2 + \frac{p_0^2}{2g} = 0 \]  

(14)

where $h_M$ is the height corresponding to the maximum value of topography function $z_M$ and can be estimated from the Bernoulli equation (15) which is the general solution for one-dimensional and steady state situations.

\[ \frac{p_0^2}{2gh(x)^2} + h(x) + z(x) = Cst \]  

(15)
In transcritical flow, regime is subcritical in the zone upwind the bump and supercritical downwind the top of the bump. To attain this flow behaviour, initial condition of fluid at rest is fixed, this time with a smaller height:

$$h(x) + z(x) = 0.66 \quad \forall x \in [0, L]$$  \hspace{1cm} (16)

with $h(0)u(0) = p_0 = 1.53$ and $h(L) = 0.66$ as boundary conditions. Due to the use of a transient scheme to obtain a steady solution, condition of outlet known-height is applied while regime is subcritical. Once supercritical regime is observed that condition is not used given that its value does not correspond to the outlet value at supercritical regime.

As shown in Figure 3, numerical solution for velocity and height presents oscillations around the analytical solution in the subcritical flow region, i.e. upwind the bump. However, the global error is still good enough considering an excellent behaviour in the supercritical flow region and the proper representation of transition zone.

Although the numerical method diverges for high shape parameter values and coarse nodal distributions, the rate of convergence of the method is similar to the subcritical case. Smallest errors are found when $c = 0.1$ and the solution diverges for all nodal distributions when $c = 1.0$. Also, the behaviour obtained for the error again agrees conventional trend of direct RBF collocation schemes when using MQ functions. Graphics of error versus nodal spacing and shape parameter are not shown for brevity.

4.2 Transient 1-D problem: Dam-break

Several analytical solutions of SWEs have been obtained for the one-dimensional dam-break situation which is a classical Riemann problem. In this case, the height function at initial time is a step. The solution consists in finding the position of advance front (water surface) starting in an instantaneous dam-break in a flat topography, i.e. $z(x) = 0$. With
the aim of validating temporal discretisation scheme, the dam-break problem without
friction or Ritter problem is solved.

Ritter’s solution describes the movement of an ideal advance front, i.e. without consid-
ering friction term \((C_f = 0)\), immediately after dam-break along a dry domain \((h = 0)\).
In this case, initial condition is given by the following step function:

\[
h(x) = \begin{cases} h_l & 0 \leq x \leq x_0 \\ 0 & x_0 < x \leq L \end{cases}
\]  

(17)

where \(h_l\) and \(x_0\) are reservoir height and dam position, respectively. Initial value for
velocity is \(u(x) = 0\) or fluid at rest condition. Height analytical solution is a parabola
connecting the first zone of height \(h_l\) and zero velocity \(0 < x < x_A\), with \(x_A\) given by
equation (18), to the third zone of zero height and zero velocity \(x_B < x < L\), where \(x_B\)
can be obtained from equation (19).

\[
x_A(t) = x_0 - t\sqrt{(gh_l)}
\]  

(18)

\[
x_B(t) = x_0 + 2t\sqrt{(gh_l)}
\]  

(19)

The simulation time is fixed to be less than the time required for the advance front
to reach the boundary \(x = L\). Therefore, height boundary conditions are fixed as the
respective height values. Solutions are obtained with \(h_l = 0.0005\), \(x_0 = 5\), \(L = 10\) and,
for numerical parameters, \(c = 0.1\), \(\Delta t = 0.1\), \(N = 321\).

The present problem is an extreme situation which shows capability and limiting as-
pects of the spatial and temporal discretisation schemes implemented. Although notorious
differences are obtained between numerical and analytical solutions for water height and
velocity (Figure 4), the developed method is able to reproduce with good accuracy water
surface position immediately downwind of the advance front located at \(x_A\). The above
is enough evidence of the scheme capability to accurately solve practical problems where
temporal changes of dependent variables are smaller. The big error obtained around
points \(x_A\) and \(x_B\) is due to the approximation order of the discretisation schemes em-
ployed. Hence, numerical results could be improved by implementing a temporal advance
scheme of greater order or by applying an upwinding strategy to the local RBF collocation
method. However, the aforementioned developments are out of the reach of the present
work, which is focused in an alternative methodology for using RBF collocation method
for shallow water problems.
4.3 Transient 2-D problem

Different two-dimensional analytical solutions for the SWEs were obtained by Thacker [9]. Here, the radially symmetrical oscillating paraboloid problem, as presented in [2], is solved by using the meshless scheme developed. The topography is a paraboloid defined by the following function:

\[ z(r) = -h_0 \left( 1 - \frac{r^2}{a^2} \right) \]  

(22)

where \( h_0 \) is the depth at central point of the domain when \( H = 0 \) (zero elevation), \( a \) is the distance between central point and zero elevation circle in the shoreline and \( r \) is the Euclidian distance from central point to any point \((x, y)\) in the domain \([0, L] \times [0, L]\) given by:

\[ r = \sqrt{\left( x - \frac{L}{2} \right)^2 + \left( y - \frac{L}{2} \right)^2} \]  

(23)

The periodic solution for the frictionless flow case, is given by:

\[ h(r, t) = h_0 \left[ \frac{\sqrt{1 - A^2}}{1 - A \cos \omega t} - 1 - \frac{r^2}{a^2} \left( \frac{1 - A^2}{(1 - A \cos \omega t)^2} - 1 \right) \right] - z(r) \]  

(24)

\[ u(x, y, t) = \frac{1}{1 - A \cos \omega t} \left[ \frac{1}{2} \omega \left( x - \frac{L}{2} \right) A \sin \omega t \right] \]  

(25)

\[ v(x, y, t) = \frac{1}{1 - A \cos \omega t} \left[ \frac{1}{2} \omega \left( y - \frac{L}{2} \right) A \sin \omega t \right] \]  

(26)

where \( \omega = \sqrt{8gh_0}/a \) is the frequency, \( r_0 \) is the distance between central point and the initial position of shoreline and \( A = (a^2 - r_0^2)/(a^2 + r_0^2) \). In order to estimate boundary
conditions from analytical solution, the following values of solution parameters are considered: \( a = 3, \ r_0 = 4, \ h_0 = 1 \) and \( L = 4 \). Transient numerical solution is attained for half oscillation period given by \( T = \frac{\pi}{\omega} \) and by fixing the following numerical parameters: \( c = 10, \ n = 8 \) and \( \Delta t = 0.01 \). Nodal distribution consists in a grid of \( N = 21 \times 21 \) points equally spaced throughout the square domain.

The numerical solution obtained for height along line \( y = x \) are presented in Figure 5 for different times, with \( r^* = \sqrt{x^2 + y^2} \). The solution is in good agreement with analytical function for the period of simulation. Nevertheless, error increases as time passes but given the oscillatory nature of the problem the error reduces once an oscillation is completed as shown in right part of Figure 5 where relative \( L_2 \)-norm error of height are presented for a greater time of simulation (\( t = 4s \)).

Although it is not shown for brevity, the two components of velocity exhibit similar behaviour than height but relative error are higher towards boundary at \( t = 1.0 \) due to the presence of a wet-dry interphase. As it happens in the dam-break problem, wet-dry transitions modelling affects the accuracy of the method due to the ill-conditioning of Jacobian matrix when \( h \) is close to zero. However, the numerical results attained for the solution of this problems shows applicability of the implemented method to solve two-dimensional flow situations.

5 CONCLUSIONS

A meshless method for the solution of two-dimensional Shallow Water Equations (SWEs) have been developed by implementing a local direct RBF collocation scheme for spatial discretisation, a first-order implicit scheme for temporal discretisation and the Newton-Rapshon method for the solution of the resulting non-linear equation system.
The method is verified by analysing the error between numerical and analytical solutions for two problems: steady subcritical and transcritical flow over a bump and frictionless dam-break with dry bed. Additionally, a transient two-dimensional problem with analytical solution is accurately simulated showing the capability of the method to solve $2-D$ flow situations. Although numerical results can be improved by implementing high-order temporal schemes and upwinding temporal discretisation, it is shown that direct RBF collocation is a proper alternative for the solution of SWEs.

REFERENCES


ESTIMATES- AND CORRECTOR-BASED MESH ADAPTATION

GAUTIER BRÈTHES\textsuperscript{1}, ADRIEN LOSEILLE\textsuperscript{2}, FRÉDÉRIC ALAUZET\textsuperscript{3} AND ALAIN DERVIEUX\textsuperscript{4}

\textsuperscript{1}INRIA  
2004 Route des lucioles, F-06902 Sophia Antipolis  
Gautier.Brethes@inria.fr  

\textsuperscript{2}INRIA  
BP 105, Domaine de Voluceau, F-78153 Le Chesnay  
Adrien.Loseille@inria.fr  
https://www-roc.inria.fr/gamma/gamma/Membres/CIPD/Adrien.Loseille/index.php

\textsuperscript{3}INRIA  
BP 105, Domaine de Voluceau, F-78153 Le Chesnay  
Frederic.Alauzet@inria.fr  
https://www-roc.inria.fr/gamma/gamma/Membres/CIPD/Frederic.Alauzet/index.en.html

\textsuperscript{4}INRIA  
2004 Route des lucioles, F-06902 Sophia Antipolis  
Alain.Dervieux@inria.fr  
http://www-sop.inria.fr/members/Alain.Dervieux/Alain_Dervieux-french.html

**Key words:** Poisson problem, Compressible flow, goal-oriented mesh adaptation, anisotropic mesh adaptation, adjoint, metric

**Abstract.** We present a novel formulation for the mesh adaptation of the approximation of a PDE. The proposed formulation extends the goal-oriented formulation, since it is equation-based and uses an adjoint. At the same time, it supersedes it as a solution-convergent method. Indeed, goal-oriented methods rely on the reduction of the error in evaluating a chosen scalar output with the consequence that as mesh size is increased (more degrees of freedom) only this output is proven to tend to its continuous analog, while the solution field itself may not converge. A remarkable throughput of goal-oriented metric-based adaptation is the mathematical formulation of the mesh adaptation problem under the form of the optimization, in the well-identified set of metrics, of a well-defined functional. In the new proposed formulation, we amplify this advantage. We search, in the same well-identified set of metrics, the minimum of a norm of the approximation error. The norm is prescribed by the user and the method allows addressing the case of multi-objective adaptation, like, for example in aerodynamics, adapting the mesh for drag, lift,
moment in one shot. In this work we consider the basic linear finite-element approximation and restrict our study to $L^2$ norm in order to enjoy second-order convergence. Numerical examples for the 2D Poisson problem and for 3D Euler flows are computed.

1 INTRODUCTION

Most simulation processes start with a smart Partial Differential Equation (PDE) and finish with a step of discretisation and approximate solution. The deviation between the continuous PDE solution and its approximation on a mesh is one of the most imperfect step of the simulation since the mesh is not chosen on completely rigorous bases. Now the engineer knows well what he/she expects: typically an approximation error sufficiently smaller than some level (frequently dimensioned by the model error). The approximation step is rather well specified: find the mesh for the given error level. Therefore, finding the best mesh would simply be a minimisation problem (looking for the mesh of specified total size, which makes the error minimum). However, a mesh does not lie in a suitable space for minimisation: a mesh is a combination of coordinates and of a kind of graph. Thus going towards the discrete world is a simplification for the unknown field, and an enormous complexification in finding a new unknown, the mesh. With the arising of Hessian-based mesh adaptation methods, e.g. [6, 5], it appeared that an anisotropic mesh can be parameterized by a continuous field defined on the computational domain, the metric. The idea of continuous metric or continuous mesh has been further formalised in [9, 10]. It remains to decide which error functional will be minimised with respect to the metric. Again, this functional should be continuous.

In Hessian-based method, the P1-interpolation error is transformed in terms of a product of continuous Hessian by ingredients from the metric (see [14] for details). The optimal continuous metric is then uniquely defined in terms of the Hessian, via the continuous metric optimality conditions. Only at the end of the process, these optimality conditions are discretised to produce an approximate optimal metric. This approach has permitted powerful anisotropic unstructured mesh adaptations.

Hessian-based method do not take into account the PDE. Consequently, it can miss important features in the solution, in particular for advective dominated PDE. In contrast, goal-oriented formulations consider a scalar output error which is expressed in terms of the approximation global error. The approximation global error is essentially solution of the linearised PDE with a local error as RHS. This results in the use of an adjoint state. In [8], the minimisation problem is expressed in terms of interpolation errors weighted by functions of the adjoint. Again, the minimum of the continuous scalar output error is uniquely defined. Again, the continuous optimality system is at last discretised.

In the new norm-oriented formulation proposed in this paper, the user can prescribe a norm or a semi-norm $|u - u_h|$ of the error, in order to minimise it with respect to the mesh. As a typical example of semi-norm, this can be the sum of square deviations on
particular outputs. Let us take an example in aerodynamics. The semi-norm \( |u - u_h| \equiv |C_l(u) - C_l(u_h)|^2 + |C_d(u) - C_d(u_h)|^2 + |C_m(u) - C_m(u_h)|^2 \) will account for minimizing the errors on lift, drag, moment measured from flow solution \( u_h \) with respect to the mesh. The proposed method will ultimately address this kind of semi-norm, assuming that, as for the goal-oriented method, the issue of a non-admissible norm is solved. As for the goal-oriented method, the proposed method takes into account the PDE features and, in case where a norm is prescribed, it produces an approximate solution field which does converge to the exact one in this norm. In this paper, the method is first demonstrated with the usual linear finite-element method in 2D. This approximation is first-order accurate for \( H^1 \) norm, but second-order accurate for \( L^2 \) norm, which we shall consider here. The method relies on the use of a corrector field, and on an \textit{a priori} error estimate from which is extracted the asymptotically largest terms of the local error.

The derivation of a corrector is proposed in Sec. 2. Next two sections are devoted to the three identified mesh adaptation formulations: a Hessian-based, minimizing an interpolation error in Sec. 3, a goal-oriented formulation and our proposal for a norm-oriented in Sec. 4. Sec. 5 is devoted to a numerical comparison between the two field-convergent formulations, viz. Hessian-based and norm-oriented for an elliptic PDE. Sec. 6 addresses the Euler PDE of aerodynamics.

2 FINER-GRID CORRECTOR FOR A GENERIC PDE

We consider a linear PDE denoted \( Au = f \) and a second-order accurate discretization of it, \( A_h u_h = f_h \). Let us assume the problem is smooth and that the approximation is in its asymptotic mesh convergence phase for the mesh \( \Omega_h \) under study, of size \( h \). Then this will be also true for a strictly two-times finer embedding mesh \( \Omega_{h/2} \). We would have:

\[
 u_h = A_h^{-1} f_h , \quad u_{h/2} = A_{h/2}^{-1} f_{h/2} \quad \Rightarrow \quad u - u_{h/2} \approx \frac{1}{4} (u - u_h)
\]  

(1)

where \( u_h \) and \( u_{h/2} \) are respectively the solutions on \( \Omega_h \) and \( \Omega_{h/2} \). We have also \( \Pi_h u - \Pi_{h/2} u_{h/2} \approx \frac{1}{4} (\Pi_h u - u_h) \). This motivates the definition of a finer-grid Defect-Correction (DC) corrector as follows:

\[
 A_h \tilde{u}'_{DC} = \frac{4}{3} R_{h/2 \rightarrow h} (A_{h/2} P_{h \rightarrow h/2} u_h - f_{h/2})
\]

(2)

where the residual transfer \( R_{h/2 \rightarrow h} \) accumulates on coarse grid vertices the values at fine vertices in neighboring coarse elements multiplied with barycentric weights, and \( P_{h \rightarrow h/2} \) linearly interpolates coarse values on fine mesh. In the case of local singularities, statement (1) is not true for uniform meshes, but we have some hints that it holds almost everywhere for a sequence of adapted meshes, according to [12]. The DC corrector \( \tilde{u}'_{DC} \) approximates \( \Pi_h u - u_h \) instead of \( u - u_h \) and can be corrected as the previous one:

\[
 u'_{DC} = \tilde{u}'_{DC} - (\pi_h u_h - u_h).
\]

(3)
This field will play a key role in the norm-oriented mesh adaptation introduced in the sequel.

3 INTERPOLATION ERROR OPTIMIZATION

3.1 Mesh parametrization

We propose to work in the continuous mesh framework, introduced in [9, 10]. The main idea of this framework is to model discrete meshes by continuous Riemannian metric fields. It allows us to define the adaptation problem as a differentiable optimization problem, i.e., to apply on the class continuous metrics a calculus of variations which cannot be applied on the class of discrete meshes. This framework lies in the class of metric-based methods.

A continuous mesh \( M \) of the computational domain \( \Omega \) is identified to a Riemannian metric field \( M = (M(x))_{x \in \Omega} \) where \( M(x) \) is a symmetric \( 3 \times 3 \) matrix. We define the total number of vertices of \( M \) as:

\[
C(M) = \int_{\Omega} \sqrt{\det(M(x))} \, dx.
\]

Given a continuous mesh \( M \), we shall say that a discrete mesh \( H \) of the same domain \( \Omega \) is a unit mesh with respect to \( M \), if each triangle \( K \in H \), defined by its list of edges \((a_i b_i)_{i=1...3}\), verifies:

\[
\forall i \in [1,3], \quad \int_0^1 \sqrt{a_i b_i \cdot M(a_i + t a_i b_i) \cdot a_i b_i} \, dt \in \left[1, \sqrt{2}\right].
\]

The rest of the paper will try to find the best metric \( M \) from various error analyses.

3.2 Interpolation-based optimal metric

Let \( u \) be any smooth enough function defined on \( \Omega \). Let \( M \) be a mesh/metric of \( \Omega \). We consider only meshes \( M \) involving enough nodes for justifying the replacement of the complete error by its main asymptotic part. The \( P^1 \) interpolation error \( |\Pi_M u - u| \) can be approximated in terms of second derivatives of \( u \) and of the metric \( M \) by the continuous interpolation error:

\[
|\Pi_M u - u| \approx |u - \pi_M u| \quad \text{with} \quad |u - \pi_M u|(x) = \frac{1}{10} \text{trace}(M^{-\frac{1}{2}}(x) |H_u(x)| M^{-\frac{1}{2}}(x))
\]

(4)

where \( |H_u| \) is deduced from \( H_u \) by taking the absolute values of its eigenvalues. Starting from:

\[
||u - \pi_M u||_{L^p(\Omega_h)} = \left( \int_{\Omega} \left( \text{trace}(M^{-\frac{1}{2}}(x) |H_u(x)| M^{-\frac{1}{2}}(x)) \right)^p \, dx \right)^{\frac{1}{p}}
\]

(5)

we define as optimal metric the one which minimizes the right hand side under the constraint of a total number of vertices equal to a parameter \( N \). After solving analytically
this optimization problem, we get the unique optimal \((\mathcal{M}_{L^p}(x))_{x \in \Omega}\) as:

\[
\mathcal{M}_{L^p} = K_p(H_u) = D_{L^p}(\text{det } |H_u|)^{\frac{1}{2p+2}} |H_u| \quad \text{and} \quad D_{L^p} = N \left( \int_{\Omega} (\text{det } |H_u|)^{\frac{p}{2p+2}} \right)^{-1},
\]

where \(D_{L^p}\) is a global normalization term set to obtain a continuous mesh with complexity \(N\) and \((\text{det } |H_u|)^{\frac{1}{2p+2}}\) is a local normalization term accounting for the sensitivity of the \(L^p\) norm. In the case of an adaptation loop for solving a Partial Differential Equation, a continuous function \(u\) is not available, only an approximate solution \(u_M\) is available. In that case, the continuous interpolation error (4) is replaced by:

\[
|u_M - \pi_M u_M|(x) = \frac{1}{10} \text{trace}(\mathcal{M}^{-\frac{1}{2}}(x)|H_{u_M}(x)|\mathcal{M}^{-\frac{1}{2}}(x))
\]

And introducing the continuous interpolation error (7):

\[
\delta_{j, \text{goal}}(\mathcal{M}) \leq |a(\Pi_M u - u_M, u_{g,M}^*) - (g, u - \Pi_M u)|
\]

4 EQUATION-BASED ADAPTATION

4.1 Scalar output “goal-oriented” analysis

The goal-oriented analysis relies on the minimization of the error \(\delta_{j, \text{goal}}(\mathcal{M})\) committed on a scalar output \(j = (g, u)\), error which we simplify as follows:

\[
\delta_{j, \text{goal}}(\mathcal{M}) = |(g, u - u_M)| = |(g, \Pi_M u - u_M + u - \Pi_M u)|.
\]

The term \(u - \Pi_M u\), similar to the main term of the Hessian-based adaptation in Section 3.2, can be explicitly approached in the same way. The term \(\Pi_M u - u_M\) will be transformed via a discrete adjoint state \(u_{g,M}^*\) defined by:

\[
\forall \psi \in V_M, \quad a(\psi_M, u_{g,M}^*) = (\psi_M, g).
\]

Then:

\[
\delta_{j, \text{goal}}(\mathcal{M}) = |a(\Pi_M u - u_M, u_{g,M}^*) + (g, u - \Pi_M u)|
\]

And introducing the continuous interpolation error (7):

\[
\delta_{j, \text{goal}}(\mathcal{M}) \leq |a(\Pi_M u - u_M, u_{g,M}^*)| + |g| |\pi_M u_M - u_M|
\]
The first variational term can be estimated as in [1]:

$$\delta_{\text{goal}}(\mathcal{M}) \leq \int_{\Omega} \left( \frac{1}{\rho} \bar{\rho} (H(u_{g,M}^*)) + |g| |\pi_{\mathcal{M}} u_{\mathcal{M}} - u_{\mathcal{M}}| + |u_{g,M}^*| |\pi_{\mathcal{M}} f - f| \right) \, d\Omega.$$ 

Where we have introduced the discrete extension of the interpolation error. It is then reasonable to try to minimize the RHS of this inequality instead of the LHS. But this involves still some difficulty due to the dependancy of adjoint state $u_{g,M}^*$ with respect to $\mathcal{M}$. We shall further simplify our functional by freezing, during a part of the algorithm, the adjoint state. The idea is that, when we change the parameter $\mathcal{M}$, $u_{g,M}^*$ is close to its (non-zero) continuous limit and is not much affected, in contrast to the interpolation errors $|\pi_{\mathcal{M}} u_{\mathcal{M}} - u_{\mathcal{M}}|$ and $|\pi_{\mathcal{M}} f - f|$. We then consider, for a given $\mathcal{M}_0$, the following optimum problem:

$$\min_{\mathcal{M}} \int_{\Omega} \left( \frac{1}{\rho} \bar{\rho} (H(u_{g,M_0}^*)) + |g| |\pi_{\mathcal{M}} u_{\mathcal{M}} - u_{\mathcal{M}}| + |u_{g,M_0}^*| |\pi_{\mathcal{M}} f - f| \right) \, d\Omega.$$ 

This will produce an optimum:

$$\mathcal{M}_{\text{opt},\mathcal{M}_0} = \arg \min_{\mathcal{M}} tr(\mathcal{M}^{-1/2}) \left( \frac{1}{\rho} \bar{\rho} (H(u_{g,M_0}^*)) + |g| |H_u| + |u_{g,M_0}^*| |H_f| \right) \mathcal{M}^{-1/2}.$$

Observing that in the integrand

$$H_{\text{goal},0} = \frac{1}{\rho} \bar{\rho} (H(u_{g,M_0}^*)) + |g| |H_u| + |u_{g,M_0}^*| |H_f|$$

is a positive symmetric matrix, we can apply the above calculus of variation and get:

$$\mathcal{M}_{\text{opt},\mathcal{M}_0} = K_1( \frac{1}{\rho} \bar{\rho} (H(u_{g,M_0}^*)) + |g| |H_u| + |u_{g,M_0}^*| |H_f|).$$

This solution can then be introduced in a fixed-point loop and will produce:

$$\mathcal{M}_{\text{opt},\text{goal}} = K_1( \frac{1}{\rho} \bar{\rho} (H(u_{g,M_{\text{opt},\text{goal}}}^*)) + |g| |H_u| + |u_{g,M_{\text{opt},\text{goal}}}^*| |H_f|).$$

Let us precise how the discrete algorithm is organised:

**Step 1:** compute the discrete state $u_{\mathcal{M}}$ on mesh $\mathcal{H}_{\mathcal{M}}$,

**Step 2:** compute the discrete adjoint state $W_{\mathcal{M}}^*$,

**Step 3:** compute optimal metric $\mathcal{M}_{\text{opt},\mathcal{M}}$,

**Step 4:** $\mathcal{M} = \mathcal{M}_{\text{opt},\mathcal{M}}$, generate $\mathcal{H}_{\mathcal{M}}$ and go to 1, until convergence.

The adaptation of this process to the Euler model of Gas Dynamics is studied in [11] for the steady case and in [2] for the unsteady case.
4.2 Norm-based functional

We are now interested by the minimization of
\[ \delta_j(\mathcal{M}) = ||u - u_\mathcal{M}||^2_{L^2(\Omega)} \]
with respect to the mesh \( \mathcal{M} \). Introducing \( u'_{DC} \) from (3) gives:
\[ \delta j(\mathcal{M}) \approx (u'_{DC}, u - u_\mathcal{M}). \] 
(10)

Let us define the discrete adjoint state \( u^*_\mathcal{M} \):
\[ \forall \psi \in V_\mathcal{M}, \quad a(\psi_\mathcal{M}, u^*_\mathcal{M}) = (\psi_\mathcal{M}, u'_{DC}). \]
(11)

Then, similarly to Section 4.1, we have to solve the following optimum problem.
\[ \min_{\mathcal{M}} \int_\Omega \left( \left[ \frac{1}{\rho} \bar{\rho}(H(u^*_\mathcal{M})) + |u'_{DC}| \right] \pi_\mathcal{M} u_\mathcal{M} - u_\mathcal{M} \right) + \left| u^*_\mathcal{M} \right| \pi_\mathcal{M} f - f \right) \] \( d\Omega \).

Exactly as for Section 4.1, we freeze the dependancy of the adjoint state.
\[ \min_{\mathcal{M}} \int_\Omega \left( \left[ \frac{1}{\rho} \bar{\rho}(H(u^*_0)) + |u'_{DC}| \right] \pi_\mathcal{M} u_\mathcal{M} - u_\mathcal{M} \right) + \left| u^*_0 \right| \pi_\mathcal{M} f - f \right) \] \( d\Omega \).

In order to get the final norm-oriented optimum \( \mathcal{M}_{opt,norm} \) we shall:

**Step 1**: first solve the linearised corrector system (3) in order to get \( u'_{DC} \),

**Step 2**: then solve the adjoint system:
\[ a(\psi, u^*_\mathcal{M}) = (u'_{DC}, \psi) \]
(12)

**Step 3**: finally put:
\[ \mathcal{M}^{(\alpha+1)} = \mathcal{K}_1\left( \left| u'_{DC} \right| + \frac{1}{\rho} \bar{\rho}H(u^*_{prio}) \right| H_u^\alpha \right) + \left| u^*_{prio} \right| \right) \] \( d\Omega \).

the three-step process being re-iterated until we get a fixed point \( \mathcal{M}_{opt,norm} = \mathcal{M}^{(\infty)} \).

5 A NUMERICAL ILLUSTRATION

We restrict our study to a benchmark of two-dimensional Poisson problems. We conjecture that the two following mesh adaptation methods produce \( L^2 \) convergent solutions to continuous. The first method, the Hessian-based method (with \( p = 2 \)), is just heuristically relying on usual finite-element estimates. The second method, our novel norm-oriented method, is directly built on the minimisation of the \( L^2 \) error norm. We do not consider goal-oriented applications, for which examples of computations can be found in [11] and [2].
5.1 Numerical features

In [4], a mesh-adaptative full-multigrid (FMG) algorithm relying on the Hessian-based adaptation criterion is designed. We first describe in short this algorithm for the Hessian-based option. A sequence of numbers \( N_k \) of vertices is specified, from a coarse mesh to finer one \( N_0 = N, N_1 = 4N, N_2 = 16N, N_3 = 64N, \ldots \). For each mesh size \( N_k \), a sequence of adapted meshes of size \( N_k \) is built by iterating the following loop:

1. computing a solution,
2. computing the optimal metric,
3. building the adapted mesh.

In (1), a multi-grid V-cycle is applied to a sufficient convergence. In (2), approximations of the Hessians are performed as in [11]. When changing of mesh, an interpolation is applied in order to enjoy a good initial condition. About 4 adaptation iterations are applied at each mesh fineness \( N_k \).

The extension of the above loop to norm-oriented adaptation consists in replacing the single Hessian evaluation by:
- the computation of the corrector, using MG and the best available (interpolated to current mesh) previous evaluation,
- the computation of the adjoint, using MG and also the best available (interpolated to current mesh) previous evaluation,
- the evaluation of (13).

5.2 APPLICATION TO A 2D BOUNDARY LAYER

This test case is taken from [7]. We solve the Poisson problem \(-\Delta u = f \) in \([0, 1] \times [0, 1] \) with Dirichlet boundary conditions and a right-hand side \( f \) chosen for having:

\[
    u(x, y) = [1 - e^{-\alpha x} - (1 - e^{-\alpha})x]4y(1 - y).
\]

The coefficient \( \alpha \) is chosen equal to 100. The graph of the solution is depicted in Figure 1a. In Figure 1b, we show a set of FMG calculations for the considered test case. The number of vertices of the successive meshes is supported by the horizontal axis, from 120 vertices to 30,000 vertices. The vertical axis gives the \( L^2 \)-norm of the approximation error \( |u - u_h|_{L^2} \) obtained on the mesh. Its variation with respect to number of vertices is compared in Figure 1b for the three following algorithms: (a) the uniform-mesh FMG, and (b) the Hessian-based adaptive FMG, and (c) the norm-oriented adaptive FMG. We observe that both adaptation methods carry an important improvement with respect to uniform-grid FMG (25921 vertices on finest mesh). For essentially the same number of vertices (32318), the Hessian option gives an error divided by 47. The norm-oriented option appears as better, with an error divided by 208 with 29485 vertices.
Figure 1: Fully 2D Boundary layer test case: (a) sketch of the solution, (b) convergence of the error norm $|u - u_h|_{L^2}$ as a function of number of vertices in the mesh for (+) non-adaptative FMG, (×) Hessian-based adaptative FMG, (★) norm-oriented adaptative FMG.

6 EULER FLOW

6.1 Methods

The above method has been extended to Euler flow adaptation. Let us denote $\Psi(W) = 0$ the steady Euler equations where $W = \{\rho, \rho u, \rho E\}$ is the set of conservation variables. Let $\Psi_h(W_h) = 0$ be its discretization by a vertex-centered second-order upwind scheme. The DC evaluation of the corrector writes:

$$\frac{\partial \Psi_h}{\partial W} \tilde{W}_{DC} = \frac{4}{3} R_{h/2-h} \Psi_{h/2}(P_{h-h/2} W_h), \quad W_{DC}' = \tilde{W}_{DC}' - (\pi_h W_h - W_h). \quad (14)$$

Then

$$\frac{\partial \Psi_h}{\partial W} g_{DC}' = W_{DC}'. \quad (15)$$

In practice, a nonlinear version of (15) is used. The rest of the algorithm is very similar to a goal-oriented algorithm, for which we follow the lines of [11].

6.2 An example

We consider the geometry provided for the 1st AIAA CFD High Lift Prediction Workshop (Configuration 1). We consider an inflow at Mach 0.2 with an angle of attack of 13 degrees. Three adaptation strategies are compared: the first one controls the interpolation error on the density, velocity and pressure in $L^1$ norm, the second controls the interpolation error on the Mach number while the third one is based on the norm-oriented approach and controls the norm of the approximation error $||W - W_h||_{L^2}$. For each case, five adaptations at fixed complexity are performed for a total of 15 adaptations with the following complexities: [160 000, 320 000, 640 000]. This choice leads to final meshes having around 1 million vertices. The residual for the flow solver convergence is set to $10^{-9}$. 

for each case. The generation of the anisotropic meshes is done with the local remeshing strategy of [13]. The surface meshes and the velocity iso-lines are depicted in Figure 2. Depending on the adaptation strategy, completely different flow fields are observed. The adaptation on the Mach number reveals strong shear layers at the wing tip that are not present in the norm oriented approach. On the contrary, recirculating flows are observed on the norm oriented approach while not being observed on the Mach number adaptation. For each case, the wakes have different features. Note that the accuracy near the body is not equivalent. For the $L^1$ norm adaptation error and norm oriented approaches, the far-field and inflow are much more refined than in the Mach number adaptation. This leads to unresolved phenomena for the final considered complexity. This example illustrates the need to control the whole flow field. Indeed, if the adaptation on the Mach number can provide a second-order convergent field, there is no guarantee on the other fields (density, pressure, velocity,...). In addition, the adaptation with the norm-oriented approach tends to increase the refinement also at the inflow boundary condition and also at the far-field although the interpolation error (on all variables) is negligible in these areas. Consequently, it seems of main interest to control all the sources of error, especially when the final intent is to certify a flow simulation.

7 CONCLUSION

The norm-oriented mesh adaptation method is an answer to a well-formulated problem: we choose an error norm and prescribe a number of nodes and we have to find the mesh giving the smallest approximation error in that norm. The norm-oriented mesh adaptation method transforms the problem into an optimization problem which is mathematically well-posed. For this, a Defect-Correction corrector is built from a finer-mesh defect correction principle. The norm-oriented method is presented as a natural extension of the goal-oriented method which, in our formulation, is itself a natural extension of the Hessian-based method. More precisely, while the Hessian-based method solves only the EDP under study, the goal-oriented method also solves an adjoint system (with linearised operator, transposed). The norm oriented solves three systems, a corrector (linearised system with an adhoc RHS), an adjoint (linearised and transposed, with the corrector as RHS), and the EDP itself. The three algorithms have in common an anisotropic a priori error analysis and a metric-based mesh parametrisation. The Hessian-based method produces convergent solution fields but does not take into account the precise equation and discretization. The goal-oriented method takes into account equation and discretization, but is too focused on a particular output and does not produce convergent solution fields. The norm-oriented method has the advantages of both. For elliptic problems, the Hessian-based approach is nearly optimal as suggested by finite-element estimates. However the presented comparisons seem to indicate that the novel method carries a good improvement. We have also proposed a preliminary application to an inviscid compressible flow. New computations will be shown during the conference.
Figure 2: Surface mesh and velocity iso-values when controlling the sum of the $L^1$ norm of the interpolation error on the density, velocity and pressure (top), the Mach number (middle) and the norm $\|W - Wh\|_{L^2}$ with the norm oriented approach (bottom).

8 ACKNOWLEDGEMENTS

This work has been supported by French National Research Agency (ANR) through project MAIDESC n° ANR-13-MONU-0010. The first author is supported by region PACA and Lemma Engineering company.

REFERENCES


A SECANT-STEPBACK PROCEDURE FOR THE SOLUTION OF SEVERE GEOMETRIC AND PHYSICAL NONLINEAR PROBLEMS

MAURO H. A. DE LIMA JUNIOR¹, FABRÍCIO N. CORREA¹, BRENO P. JACOB¹, BRUNO M. JACOVAZZO² AND BRUNO F. MONTEIRO²

¹ LAMCSO/COPPE/UFRJ Rio de Janeiro, RJ, Brazil
LAMCSO – Laboratory of Computational Methods and Offshore Systems, Civil Engineering Dept.
Post Graduate Institute of the Federal University of Rio de Janeiro
Av. Pedro Calmon, S/N, Cidade Universitária, Ilha do Fundão, Caixa Postal 68.506,
CEP: 21941-596, Rio de Janeiro, RJ, Brazil.
{mauro, fabricio, breno}@lamcso.coppe.ufrj.br

² POLI/UFRJ Rio de Janeiro, RJ, Brazil
Polytechnic School – Federal University of Rio de Janeiro
Av. Athos da Silveira Ramos, nº 149, CT - Bloco D, Sala D-205, Cidade Universitária, Ilha do
Fundão, Caixa Postal 68.529, CEP: 21945-970, Rio de Janeiro, RJ, Brazil.
{brunojacovazzo, bruno.monteiro}@lamcso.coppe.ufrj.br

Key words: Nonlinear Analysis, Computational Methods, Ramberg-Osgood and FEM.

Abstract. This work presents an iterative strategy for the solution of structural problems submitted to severe physical and geometrical nonlinear effects. This strategy has been originally devised to analyze installation procedures for submarine pipelines and other slender offshore lines [1]. These structural systems are characterized by severe geometric nonlinear effects due to their large displacements, and also by a markedly physical nonlinear behavior due, for instance, to the utilization of ancillary systems such as bending restrictors, or also to the use of the Ramberg-Osgood model [2] for the representation of the nonlinear elastic and elastoplastic behavior in pipelines.

The proposed strategy is characterized by the association of the classical Newton-Raphson iterative method (evaluating the geometrical nonlinear behavior of the global effective matrices of the finite elements) with a so-called “Secant-Stepback” procedure that is performed at the local level of each element to solve its physical nonlinear behavior. Nonlinear solution strategies based on this concept of secant stiffness have been presented in the literature (see for instance [3]), but in some situations their application to such slender offshore systems with severe physical and geometrical nonlinear effects has faced convergence problems, thus failing to efficiently lead to the desired solution.

In this context, the “Secant-Stepback” procedure is intended to address any kind of physical nonlinear behavior of the material, but in this work its use is exemplified in terms of a moment-curvature curve (in the case of ancillary systems such as bending restrictors, with hardening behavior), and in terms of the Ramberg-Osgood formulation representing elastoplastic materials with softening.
1 INTRODUCTION

For strongly nonlinear problems such as flexible lines with bending restrictors, it is known that the Newton-Raphson methods (NRP and NRM) require a very large number of iterations to reach convergence, and depending on the degree of nonlinearity there are situations of divergence, requiring the increase of number of steps to solve the problem.

Such behavior is associated with the sequential step-by-step iterative process of balance of force-displacement that uses the inner force in the nonlinear curve (NL) and corresponding tangent stiffness of the previous step to solve the displacements corresponding to the current step of equilibrium. In strongly nonlinear problems type of search may result, in the iterative process, very high force-displacement unbalances, which when not present divergences require various iterations to achieve convergence. Even in methods presented most recently, as the Secant Method (MS) developed by WEN and ZENG in 2010 [3], this phenomenon can still be observed.

In this sense, it is the differential of Secant-Stepback Method (STBK-S), described to follow, designed to solve physical nonlinear problems with operations at the level of each element, taking advantage of N-L curve preset (for example, by formulating Ramberg-Osgood, or by a curve completely defined by points curvature-moment). Therefore, it is possible to know the displacement that would occur due to external force acting on the element. Once known this displacement, secant stiffness can be easily obtained, as will be presented in more detail below.

2 DESCRIPTION OF THE METHOD

Secant-Stepback Method consists an iterative process for solving nonlinear physical problems at the local level of element by successive approximations, in order to conduct the numeric response a minimum error according to a convergence criterion.

Subscript index \( e \) in following parameters represents procedures at the local level of elements.

From the nonlinear static equilibrium equation, and considering that the external load \( \{ \Delta R \}_e \) can vary in local axis for each iteration, an iterative process for the convergence of the solution method is required. Eq. (1) shows such process.

\[
\begin{align*}
\{ U \}_e^{(i)} & = \{ F \}_e^{(i)} - \{ \Delta U \}_e^{(i)} \\
& = \{ K_{sec} \}_e^{(i-1)} \{ \Delta U \}_e^{(i)} + \{ \Delta U \}_e^{(i)} \\
& = \{ K_{sec} \}_e^{(i-1)} \{ \Delta U \}_e^{(i)} + \{ F \}_e^{(i)} - \{ F \}_e^{(i-1)} \\
& = \{ K_{sec} \}_e^{(i-1)} \{ \Delta U \}_e^{(i)} + \{ \Delta F \}_e^{(i)}
\end{align*}
\]

Where,
- \( \{ K_{sec} \}_e^{(i-1)} \) is the secant stiffness matrix.
- \( \{ \Delta U \}_e^{(i)} \) is the incremental displacement.
- \( \{ R \}_e^{(i)} \) is the external force at the current step.
- \( \{ F \}_e^{(i)} \) is the internal force at the previous time \( t \).
- \( U \) is the displacement generated by external force on nonlinear curve obtained from the external force \( \{ R \}_e^{(i)} \) and initial displacement \( \{ U \}_e^{(i)} \). Thus, the increase
of displacement of the proposed iterative procedure is calculated according to Eq. (2).

\[
\{\Delta U\}_{e}^{(i)} = i^{+\Delta t}\{U^{(i+\Delta t)}\}_{e}^{(i)} - i\{U\}_{e}^{(i)}
\]  

(2)

Next step consists to compute the secant stiffness in element by Eq. (3).

\[
i^{+\Delta t}\{K_{sec}\}_{e}^{(i-1)} = \frac{i^{+\Delta t}\{R\}_{e}^{(i)} - i\{F\}_{e}}{\{\Delta U\}_{e}^{(i)}}
\]  

(3)

Now, Eq. (4) shows the calculation of the displacement.

\[
i^{+\Delta t}\{U\}_{e}^{(i)} = i\{U\}_{e}^{(i)} + \left(i^{+\Delta t}\{K_{sec}\}_{e}^{(i-1)}\right)^{-1}\left(i^{+\Delta t}\{R\}_{e}^{(i)} - i\{F\}_{e}\right)
\]  

(4)

Error is based on the relative difference of the secant stiffness between current iteration and previous one, according to Eq. (5).

\[
i^{+\Delta t}\{e\}_{e}^{(i)} = \frac{i^{+\Delta t}\{K_{sec}\}_{e}^{(i)} - i^{+\Delta t}\{K_{sec}\}_{e}^{(i-1)}}{i^{+\Delta t}\{K_{sec}\}_{e}^{(i-1)}}
\]  

(5)

In case of divergence, the method returns to the previous step replacing the stiffness of the material by the secant stiffness evaluated by the STBK-S method, restarting the solution process based on NRS method, to obtain a new equilibrium position. Thus, the secant stiffness is reevaluated, repeating the calculations of equations from (1) to (5).

In case of convergence, the residual force between the external force and the internal force in nonlinear curve must be calculated and added as pseudo-force to next time step, as shown in Eq. (6).

\[
i^{+\Delta t}\{DF\}_{e}^{(i)} = i^{+\Delta t}\{R\}_{e}^{(i)} - i\{F^{(i^{+\Delta t}\{U\}_{e}^{(i)})}\}_{e}
\]  

(6)

Eq. (7) presents the initial conditions for each increment.

\[
i^{+\Delta t}\{U\}_{e}^{(0)} = i\{U\}_{e}, \quad i^{+\Delta t}\{K\}_{e}^{(0)} = i\{K\}_{e}, \quad K_{e} = \frac{\partial R^{(0)}}{\partial U}
\]

\[
i^{+\Delta t}\{F\}_{e}^{(0)} = i\{F\}_{e} + i\{DF\}_{e}, \quad (i = 1, 2, \ldots)
\]  

(7)

Iterative process continues until the error \(i^{+\Delta t}\{e\}_{e}^{(i)}\) is less than the prescribed tolerance.

A Schematic illustration of the Secant-Stepback iterative process for a single variable can be seen in Fig. 1.

Secant-Stepback method (STBKS) has a high rate of convergence due to the following features:

- It seeks the internal force in force vs. displacement (known) curve by the force and displacement in the past, and the future force to each generation.

- Method reports the difference between internal and external force \(\{DF\}_{e}^{(i)}\), after reaching the local convergence, which next time step comes as pseudo-force in system to be solved by Modified Newton-Raphson (NRM) globally.
Figure 1: Secant-Stepback method.

The re-evaluation of the stiffness matrix in each iteration does not require high computational cost, since the use of the secant matrix accelerates convergence process. Moreover, for a material with a strongly non-linear behavior, such as hardening materials or bending restrictors, the method reaches fully convergence.

3 CASE STUDIES AND ANALYSIS OF RESULTS

This section presents two case studies aiming illustrate the applicability, efficiency and robustness of the proposed algorithm to treat high physical nonlinearity. They consist in two structural systems modeled by finite elements and analyzed by nonlinear static and dynamic simulations. A comparison between Secant-Stepback Method (STBK-S) and Standard Newton-Raphson Method (NRS) is done.

3.1 Case Study 1: Suspended Line under Action of the Self-Weight

This example aims to verify the behavior of the algorithm based on balance of force caused by the action of self-weight, in static simulations. Will be reviewed the following situations:

i. Application of 1 load step;

ii. Application of 20 load steps;

iii. Check the number of steps needed to achieve convergence in NRM method.

3.1.1 Model Characteristic (Case Study 1)

The model consists of a flexible line in suspended catenary, completely immersed in water, with the ends positioned under the same quote, as shown in Figure 2.
Coordinates of connections and boundary conditions are presented in Tables 1 and 2, respectively. It is observed that both ends are articulated; the point A is free to move in plane of line, while the point B is fixed in all directions of translation.

Templates are composed of 200 non-linear three-dimensional beam elements, whose geometric, physical and hydrodynamic properties are presented in Tables 3 and 4.

<table>
<thead>
<tr>
<th>Geometric Properties</th>
<th>Physical Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line Length (L) 200.0 m</td>
<td>Poisson Coefficient (ν) 0.3</td>
</tr>
<tr>
<td>Outer Diameter (OD) 0.5816 m</td>
<td>Axial Stiffness (EA) 200000 kN</td>
</tr>
<tr>
<td>Internal Diameter (d) 0.4826 m</td>
<td>Weight in Air (PA) 4.06968 kN/m</td>
</tr>
<tr>
<td></td>
<td>Weight in Water (PW) 1.39842 kN/m</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hydrodynamic Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inertia Coefficient (CM) 2.0</td>
</tr>
<tr>
<td>Normal Drag Coefficient (CDN) 1.2</td>
</tr>
<tr>
<td>Add Mass Coefficient (CA) 1.0</td>
</tr>
</tbody>
</table>

Solution of the nonlinear iterative process was held by Secant-Stepback (to solve the physical nonlinear problem in element level) associated with Standard Newton-Raphson Method (NRS) (to evaluate the geometric nonlinearities in global level by means of the effective matrix of the system). The convergence parameters used in the analysis for NRS are presented in Table 5 below.

<table>
<thead>
<tr>
<th>Tolerance</th>
<th>Force</th>
<th>Displacement</th>
<th>Maximum Number of Iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.001</td>
<td>0.01</td>
<td>45</td>
</tr>
</tbody>
</table>
Nonlinear behavior of the physical material, with hardening, which comprises the line is defined by points bending moment, shown in Table 6 and Figure 3 below.

![Figure 3: Characteristic curve of material.](image)

<table>
<thead>
<tr>
<th>Curvature [m⁻¹]</th>
<th>Bending Moment [kN.m]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.040</td>
<td>196.200</td>
</tr>
<tr>
<td>0.041</td>
<td>206.010</td>
</tr>
<tr>
<td>0.042</td>
<td>225.630</td>
</tr>
<tr>
<td>0.043</td>
<td>264.870</td>
</tr>
<tr>
<td>0.044</td>
<td>343.350</td>
</tr>
<tr>
<td>0.045</td>
<td>500.310</td>
</tr>
<tr>
<td>0.046</td>
<td>814.230</td>
</tr>
<tr>
<td>0.047</td>
<td>1442.070</td>
</tr>
<tr>
<td>0.048</td>
<td>2697.750</td>
</tr>
<tr>
<td>0.049</td>
<td>5209.110</td>
</tr>
<tr>
<td>0.050</td>
<td>10231.830</td>
</tr>
<tr>
<td>0.051</td>
<td>20277.270</td>
</tr>
<tr>
<td>0.080</td>
<td>602912.790</td>
</tr>
<tr>
<td>2.000</td>
<td>77751891.990</td>
</tr>
</tbody>
</table>

3.1.2 Analysis of the Results (Case Study 1)

Results presented here refer to the static simulations of the models described in the previous item, based on 2 (two) methods of non-linear problems solution: i. Secant-Stepback (STBK-S) associated with NRS and ii. NRS. The Table 7, the following, presents a summary of the convergence process attained by methods.

<table>
<thead>
<tr>
<th>Number of Load Steps</th>
<th>Solution i) STBK-S</th>
<th>Solution ii) NRS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Solution Processing Time</td>
<td>CPU Time</td>
</tr>
<tr>
<td>1</td>
<td>Converges 2.50s</td>
<td>Diverges -</td>
</tr>
<tr>
<td>20</td>
<td>Converges 2.69s</td>
<td>Diverges form step 7</td>
</tr>
<tr>
<td>140</td>
<td>Converges 3.77s</td>
<td>Converges 27.50s</td>
</tr>
</tbody>
</table>

Analyzing convergence process summary table, Table 7, one can realize that the NRS method (ii) reached convergence only after 140 steps of load (CPU Time = 27.5s), while the STBK-S method converges in just 1 step of load (CPU Time = 2.5s). Besides STBK-S presents a solution for 1 step, the method was also evaluated with 140 steps to be compared with NRS. In this case, STBK-S presents a computational cost of just 3.77s (7 times faster than the NRS), evidencing the computational efficiency of the proposed method.
Figure 4, below, presents a comparison between deformed configurations of the line after the static balance condition. It is worth mentioning that the results concerning the NRP method for 20 steps of load presented in the following graphs refer to the last step of its convergence (step 6), that is not the final equilibrium position.

Analyzing the Figure 4 above, one can realize that the final deformed configurations (equilibrium position) are quite distant from the original configuration (in catenary), with a distance of approximately 150 m between points A’ and A, resulting in a situation of high nonlinear problem. One realizes yet that deformed configurations obtained through STBK-S method with 1 step and 20 steps are already coincident, for this reason, the results from STBK-S with 140 steps was omitted from the results. Notice also that deformed configuration presented to the NRS method, referring to step 6, is distant of presented by STBK-S method, once the NRS presented divergence from step 7. It is possible to observe that the deformed configuration corresponding to the NRS after applying of 140 steps of load matches with the results obtained by SBTK-S method with 1 step of load, signaling that the proposed method was able to handle efficiently the physical and geometrical non-linearity of the templates in static equilibrium condition.

Following Figures 5 and 6 present the envelopes of bending moment and curvature, obtained after the static equilibrium.

Analyzing the graphs above one can realize, as well as deformed in configuration, that the results obtained on the basis of the STBK-S method with one and 20 steps are coincident with those presented by the NRS method after application of 140 steps of load. This indicates that the proposed method can handle non-linear effects in static condition. Notice that the results obtained on the basis of NRS method with 20 steps of loads are distant from the obtained by SBTK-S method, as expected, since the NRS method diverged in step 7.
Figure 5: Envelope of bending moment.

Figure 6: Envelope of curvature.

Figure 7 presents the graph of curvature vs. bending moment, obtained after the static equilibrium, and also the characteristic curve of the material.

Figure 7: Graphic of bending moment vs. curvature.
Analyzing Figure 7, it can be affirmed that the nonlinear solution procedure implemented, STBK-S method, updates the rigidity of each element based on the characteristic curve of the material, since the results based on this method fully coincide with the characteristic curve of steel. It is also noticed that, the result of the NRS method considering 20 steps of load, overlap only the linear stretch of the characteristic curve, indicating that the method stopped to converge as soon as the line entered the nonlinear regime, requiring the application of 140 steps of load to reach convergence.

3.2 Case Study 2: Suspended Line under Action of the Prescribed Accelerated Motion

This example aims verify the behavior of the algorithm based on balance of force caused by the action of a prescribed accelerated movement associated with one end of the flexible line in catenary suspended.

3.2.1 Model Characteristic (Case Study 2)

This model is similar to the previous case study, with the addition of a dynamic simulation with prescribed motion with high acceleration in double back-and-forth associated with the free end A, as illustrated in Figure 8 below.

The iterative non-linear methods studied here was the same used in Case Study 1. The convergence parameters are the same as the previous example.

Dynamic analyses were performed using the integration method \( \alpha \beta \)-Newmark (implicit) with total simulation time equal to 1400s with distinct integration ranges for each solution method, see Table 8, in order to troubleshoot convergence of NRS method.
3.2.2 Analysis of Result (Case Study 2)

Figures 9 and 10, to follow, present the time series of bending moment and curvature in the central element of the line (element 100), based on the average values of moment and curvature due to application of solution methods STBK-S and NRS.

Analyzing the time series presented in Figures 9 and 10, we realize the complete juxtaposition between the results obtained through the methods STBK-S and NRS, which attests to the validity and consistency of the results of the proposed method, because it presents the same results of the NRS classic method.

Analyzing the time series presented above, we can see the emergence of regions where the moments and curvatures remain constant during dynamic simulations (minimum values: 0s-200s, 600s-800s, 1200s-1400s, and maximum values: 300s-500s, 900s-1100s), indicating the numeric stability of the response. One can also see a repetitive behavior of these values along the simulations, mainly due to the cyclical nature of the prescribed movement, which have to signal that the STBK-S algorithm does not accumulate errors in the solution process. It also observed small variations in the responses around 300s, 500s, 900s and 1100s, which are associated with the inertial effects due to the imposed motion at the connection point A.

Table 8: Numerical integration ranges.

<table>
<thead>
<tr>
<th>Solution Method</th>
<th>Integration Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>STBK-S</td>
<td>0.1s e 0.01s</td>
</tr>
<tr>
<td>NRS</td>
<td>0.01s</td>
</tr>
</tbody>
</table>

Figure 9: Time history of bending moment.

Figure 10: Time history of curvature.
Figure 11, to follow, presents a comparison between the deformed configurations of line, with basis on the solution methods STBK-S and NRS, at the point of maximum displacement of the connection point A, in average times (400s and 1000s) corresponding to the maximum values of bending moment and curvature.

Analyzing Figure 11, one can realize that deformed configurations obtained by applying both methods STBK-S and NRS are coincident. It shows that STBK-S method actually has the same results for the classic method NRS.

The Figure 12 above presents a comparison of the characteristic curve of the material and the graphics of bending moment vs. curvature in the central element of the line, based on the application of both methods of solution. One can observe the complete juxtaposition of results obtained for both methods STBK-S and NRS in relation to the material characteristic curve, indicating the STBK-S method is able to treat the non-linear physical behavior of the line based on the characteristic curve of the material without accumulating errors even under cyclic loads.

It is worth mentioning the computational cost involved in the solution process of the methods analyzed, which are presented in the Table 9, to follow.

### Table 9: Numerical integration ranges.

<table>
<thead>
<tr>
<th>Solution Methods</th>
<th>Integration Interval</th>
<th>Number of Evaluations of the Stiffness Matrix</th>
<th>Computacional Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>STBK-S</td>
<td>0.1</td>
<td>92408</td>
<td>5min 22s</td>
</tr>
<tr>
<td>STBK-S</td>
<td>0.01</td>
<td>599495</td>
<td>34min 9s</td>
</tr>
<tr>
<td>NRS</td>
<td>0.01</td>
<td>614296</td>
<td>34min 45s</td>
</tr>
</tbody>
</table>

Figure 11: Deformed configuration.  
Figure 12: Graphic of bending moment vs. curvature.
It should be noted that STBK-S method was analyzed with the integration interval 0.01s only for comparison effect of the computational cost with the NRS method, once this methods reaches the convergence only to integration interval less than or equal to 0.01s.

4 CONCLUDING REMARKS

This work presents an iterative strategy for the solution of structural systems submitted to severe nonlinear effects, both physical and geometrical. These structural systems are characterized by severe geometric nonlinear effects due to their large displacements, and also by a markedly material nonlinear behavior due, for instance, to the utilization of ancillary systems such as bending restrictors, or also to the use of the Ramberg-Osgood model for the representation of the nonlinear elastic and elastoplastic behavior in pipelines.

Proposed strategy is characterized by the association of the classical Newton-Raphson iterative method (where the operations are performed at a global level, corresponding to the global effective matrices assembled from the contributions of the finite element matrices) with a so-called “Secant-Stepback” procedure that is performed at the local level of each element. The “Secant-Stepback” procedure is intended to address the physical nonlinear behavior of the material, which is defined either in terms of a moment-curvature curve (in the case of ancillary systems such as bending restrictors, with hardening behavior), or else in terms of the Ramberg-Osgood formulation representing elastoplastic materials with softening. In either case, the material law is known a-priori, and a stepback procedure is then performed at the element level to allow the definition, at each iteration of the global Newton-Raphson procedure, of the secant stiffness. On the other hand, the geometric nonlinear behavior of the slender structure is addressed by the global NRM method.

Results of the case studies presented confirms the robustness and efficiency of STBK-S method, since it was observed complete compatibility between the results from the application of the methods STBK-S and NRS, observing a still lower computational cost associated with the STBK-S method. The results showed that the method does not accumulate errors in the iterative process of solution, being possible to simulate nonlinear behaviors with stiffening under cyclic loads.

It should be reinforced that the STBK-S method is much more robust for troubleshooting strongly non-linear, allowing the use of integration interval higher than NRS method.

REFERENCES

A VARIATIONAL FLUX RECOVERY APPROACH FOR ELASTODYNAMICS PROBLEMS WITH INTERFACES

PAVEL BOCHEV¹ AND PAUL KUBERRY¹

¹ Computational Mathematics, Sandia National Laboratories *, Mail Stop 1320, Albuquerque, New Mexico, 87185-1320, {pbboche,pakuber}@sandia.gov

Key words: Linear elastodynamics, material interface, partitioned solution.

Abstract. We present a new explicit algorithm for linear elastodynamic problems with material interfaces. The method discretizes the governing equations independently on each material subdomain and then connects them by exchanging forces and masses across the material interface. Variational flux recovery techniques provide the force and mass approximations. The new algorithm has attractive computational properties. It allows different discretizations on each material subdomain and enables partitioned solution of the discretized equations. The method passes a linear patch test and recovers the solution of a monolithic discretization of the governing equations when interface grids match.

1 INTRODUCTION

This paper focuses on the numerical solution of elastodynamic problems with interfaces. Such problems arise in multiple modeling and simulation contexts involving elastic bodies with discontinuous material properties. We present a new explicit scheme for such problems, which uses variational flux recovery techniques [2] to enable partitioned solution of the interface problem. Restriction of the governing equations to material subdomain yields boundary value problems linked through unknown interface traction. Approximation of the latter by variational recovery techniques decouples the subdomain problems.

The resulting algorithm has some attractive computational properties. It allows the use of different discretizations on each material subdomain and enables partitioned solution of the discretized equations. This makes it possible to also use the algorithm as a coupling tool for different codes operating in different material regions. The method passes a

*Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy’s National Nuclear Security Administration under contract DE-AC04-94AL85000.
linear patch test and recovers the solution of a monolithic discretization of the governing equations when interface grids match.

To present the method it suffices to consider small displacements and linear elastic models. Our main focus is on enabling explicit solution of the elastodynamic problem by solving independent problems on each material subdomain. Thus, we restrict attention to subdomain partitions with non-matching but spatially coincident interface grids. This is in contrast to many of the existing works on elliptic problems with interfaces, which focus primarily on capturing weak and strong discontinuities of the solution on unfitted grids. Such methods often combine Nitsche’s method with extended finite elements [1, 3, 4], or define a suitable modified or enriched basis set on cut elements [5, 6].

2 NOTATIONS

We consider a bounded region $\Omega \subset \mathbb{R}^d$, $d = 2, 3$ with a material interface $\sigma$. The interface splits $\Omega$ into non-overlapping subdomains $\Omega_1$ and $\Omega_2$ with Dirichlet boundaries $\Gamma_i = \partial \Omega_i / \sigma$, $i = 1, 2$. We assume that the interface unit normal $\mathbf{n}_\sigma$ coincides with the outer unit normal to $\partial \Omega_1$. The Sobolev space for the displacements on $\Omega_i$ is $H(\Omega_i)$ and $H_{\Gamma_i}(\Omega_i)$ is its subspace of functions that vanish on $\Gamma_i$. Each subdomain is endowed with a finite element partition $\Omega_i^h$. The set of all mesh vertices $\{x_{i,k}\}$ is $V(\Omega^h_i)$ and $V(\Omega^h_i)$ are the interior mesh vertices. The subdomain partitions induce finite element partitions $\sigma_i^h$ and $\sigma_i^0$ of the interface $\sigma$, which are not required to match but are assumed to be spatially coincident. The set $V(\sigma_i^h)$ contains the vertices of the interface mesh $\sigma_i^h$.

$S_i^h$ is a conforming finite element subspace of $H(\Omega_i)$, defined on the mesh $\Omega_i^h$ and equipped with a standard Lagrangian basis $\{N_{i,k}\}$. The interface part $S_i^h(\sigma)$ is the span of all basis functions associated with $V(\sigma_i^h)$ and $S_i^{h,0}$ is the space corresponding to the interior vertices $V(\bar{\Omega}_i^h)$. $S_i^{h,1}$ is a conforming subspace of $H_{\Gamma_i}(\Omega_i)$. The coefficient vector of $u_i \in S_i^h$ is $\bar{u}_i = (\bar{u}_{i,\sigma}, \bar{u}_{i,0})$ where $\bar{u}_{i,\sigma}$ and $\bar{u}_{i,0}$ are the interface and interior coefficients of $u_i$, respectively, corresponding to functions $u_{i,\sigma} \in S_i^{h,\sigma}$ and $u_{i,0} \in S_i^{h,0}$, respectively. The operator $\Pi_1 : S_i^{h,\sigma} \mapsto S_i^{h,1}$ interpolates $u_{2,\sigma} \in S_i^{h,2}$ in $S_i^{h,1}$, i.e.,

$$\Pi_1(u_{2,\sigma}) = \sum_{i \in V(\sigma_i^1)} u_{2,\sigma}(x_{1,i}) N_{1,i}(x) = \sum_{i \in V(\sigma_i^1)} \left( \sum_{k \in V(K_2 \ni x_{1,i})} (\bar{u}_{2,\sigma})_k N_{2,k}(x_{1,i}) \right) N_{1,i}(x), \quad (1)$$

where $V(K_2 \ni x_{1,i})$ are the vertices of element $K_2 \subset \sigma_i^h$ containing vertex $x_{1,i}$ from $\sigma_i^h$. The coefficient vector of $\Pi_1(u_{2,\sigma})$ is given by $P_1 \bar{u}_{2,\sigma}$ where $P_1$ is a $|V(\sigma_i^1)| \times |V(\sigma_i^2)|$ sparse matrix. The row of this matrix corresponding to vertex $x_{1,i}$ contains the values $N_{2,k}(x_{1,i})$ for $k \in V(K_2 \ni x_{1,i})$. Similar representation holds for $\Pi_2 : S_i^{h,1} \mapsto S_i^{h,2,\sigma}$.

3 Governing equations

We write the model elastodynamic problem as a pair of governing equations

$$\begin{cases}
\ddot{u}_i - \nabla \cdot \sigma(u_i) = f & \text{in } \Omega_i \times [0, T] \\
u_i = g & \text{on } \Gamma_i \times [0, T]
\end{cases}
\quad \text{and} \quad
\begin{cases}
u_i(0, x) = u_0(x) & \text{in } \Omega_i \\
\dot{u}_i(0, x) = \dot{u}_0(x) & \text{in } \Omega_i
\end{cases} \quad (2)$$

Pavel Bochev and Paul Kuberry
for displacements \( u_i(t, x) \), \( i = 1, 2 \) in \( \Omega_i \), and a pair of interface conditions
\[
\begin{align*}
  u_1(x, t) &= u_2(x, t) \quad \text{and} \quad \sigma_1(x, t) \cdot n_\sigma = \sigma_2(x, t) \cdot n_\sigma \quad \text{on} \quad \sigma \times [0, T]
\end{align*}
\]
expressing continuity of the displacement and the traction across the interface. We restrict attention to linear elastodynamic problems for which
\[
\sigma(u_i) = \lambda_i (\nabla \cdot u_i) I + 2\mu_i \varepsilon(u_i), \quad \varepsilon(u_i) = \frac{1}{2} (\nabla u_i + \nabla u_i^T),
\]
and the Lame coefficients \( \lambda_i \) and \( \mu_i \) are allowed to have a discontinuity along \( \sigma \).

4 Formulation of the method

A formal splitting of (2)–(3) into two “independent” mixed boundary value subdomain equations is the starting point in the formulation. This partitioning is formal because it imposes the unknown traction value on the interface as a Neumann boundary condition and resulting solutions satisfy a weak continuity relation in terms of an operator that is not available in closed form. Using variational flux recovery ideas we eliminate the unknown traction from the subdomain equations. In so doing we obtain two fully decoupled subdomain equations which implicitly incorporate appropriate discrete notions of the interface conditions (3).

4.1 Formal partitioning of the governing equations

Let \( u_i, i = 1, 2 \) denote the exact solutions of (2)–(3) and \( \gamma = \sigma_1(x, t) \cdot n_\sigma = \sigma_2(x, t) \cdot n_\sigma \) be the corresponding exact interface traction. If \( \gamma \) is known exactly then the displacement \( u_i \) on \( \Omega_i \) can be determined by solving the following mixed boundary value problem:
\[
\begin{align*}
  \ddot{u}_i - \nabla \cdot \sigma_i &= f \quad \text{in} \quad \Omega_i \times [0, T] \\
  u_i &= g \quad \text{on} \quad \Gamma_i \times [0, T] \quad \text{and} \quad u_i(0, x) = u_0(x) \quad \text{in} \quad \Omega_i \\
  \sigma_i(x, t) \cdot n_\sigma &= \gamma \quad \text{on} \quad \sigma \times [0, T]
\end{align*}
\]

The exact traction \( \gamma \) specifies a Neumann boundary condition on \( \sigma \), which closes the subdomain problems and makes it possible to solve them independently from each other. By the uniqueness of the solutions to (2)–(3) and (4) it follows that the solutions of the latter necessarily satisfy the first interface condition in (3), i.e., \( u_1 = u_2 \) on \( \sigma \).

The weak form of the equations in (4) are: seek \( u_i \in H(\Omega_i), i = 1, 2 \) such that
\[
\begin{align*}
  \langle \ddot{u}_1, v_1 \rangle_{\Omega_1} + \langle \sigma_1, \varepsilon(v_1) \rangle_{\Omega_1} &= \langle f, v_1 \rangle_{\Omega_1} + \langle \gamma, v_1 \rangle_{\sigma} \quad \forall v_1 \in H_{\Gamma_1}(\Omega_1) \\
  \langle \ddot{u}_2, v_2 \rangle_{\Omega_2} + \langle \sigma_2, \varepsilon(v_2) \rangle_{\Omega_2} &= \langle f, v_2 \rangle_{\Omega_2} - \langle \gamma, v_2 \rangle_{\sigma} \quad \forall v_2 \in H_{\Gamma_2}(\Omega_2)
\end{align*}
\]

Solutions of (5) necessarily satisfy the first interface condition in (3), i.e., \( u_1 = u_2 \) on \( \sigma \).
4.2 Spatial discretization

The finite element spatial discretization of (5) is to seek \( u_i \in S_i^h \times [0, T] \), which satisfies the initial and boundary conditions in (4) and is such that

\[
\begin{align*}
(u_1, v_1)_{\Omega_1} + (\sigma(u_1), \varepsilon(v_1))_{\Omega_1} &= (f, v_1)_{\Omega_1} + \langle \gamma, v_1 \rangle_{\sigma} & \forall v_1 \in S_{1,\Gamma}^h \\
(u_2, v_2)_{\Omega_2} + (\sigma(u_2), \varepsilon(v_2))_{\Omega_2} &= (f, v_2)_{\Omega_2} - \langle \gamma, v_2 \rangle_{\sigma} & \forall v_2 \in S_{2,\Gamma}^h
\end{align*}
\]  
(6)

Since in general \( \sigma_1^h \) and \( \sigma_2^h \) are non-matching finite element partitions of \( \sigma \), solutions of (6) can only satisfy a “weak” notion of displacement continuity

\[ u_{1,\sigma} = \Pi_1(u) u_{2,\sigma} \quad \text{and} \quad u_{2,\sigma} = \Pi_2(u) u_{1,\sigma}, \]  
(7)

where \( \Pi_1(u) : S_1^h \mapsto S_2^h \) and \( \Pi_2(u) : S_2^h \mapsto S_1^h \) are some unknown operators.

4.3 Elimination of the surface traction

With some abuse of notation we use \( (\cdot, \cdot)_{\Omega_k} \) to denote a discrete inner product such that \( (N_{k,i}, N_{k,j})_{\Omega_k} = 0 \) for \( i \in V(\sigma_k^h) \) and \( j \in V(\Omega_k^h) \). Then, (6) can be written in a block form corresponding to the partitioning of \( S_i^h \) into an interfacial part \( S_{i,\sigma}^h \) and a zero trace part \( S_{i,0}^h \), along with the appropriate weak continuity equation.

\[
\begin{align*}
(\hat{u}_{1,\sigma}, N_{1,i})_{\Omega_1} + (\sigma(\hat{u}_1), \varepsilon(N_{1,i}))_{\Omega_1} &= (f, N_{1,i})_{\Omega_1} + \langle \gamma, N_{1,i} \rangle_{\sigma} & \forall i \in V(\sigma_1^h) \\
(\hat{u}_{1,0}, N_{1,i})_{\Omega_1} + (\sigma(\hat{u}_1), \varepsilon(N_{1,i}))_{\Omega_1} &= (f, N_{1,i})_{\Omega_1} & \forall i \in V(\Omega_1^h) \\
u_{1,\sigma} &= \Pi_1(u) u_{2,\sigma}
\end{align*}
\]  
(8)

on the first subdomain and by

\[
\begin{align*}
(\hat{u}_{2,\sigma}, N_{2,i})_{\Omega_2} + (\sigma(\hat{u}_2), \varepsilon(N_{2,i}))_{\Omega_2} &= (f, N_{2,i})_{\Omega_2} - \langle \gamma, N_{2,i} \rangle_{\sigma} & \forall i \in V(\sigma_2^h) \\
(\hat{u}_{2,0}, N_{2,i})_{\Omega_2} + (\sigma(\hat{u}_2), \varepsilon(N_{2,i}))_{\Omega_2} &= (f, N_{2,i})_{\Omega_2} & \forall i \in V(\Omega_2^h) \\
u_{2,\sigma} &= \Pi_2(u) u_{1,\sigma}
\end{align*}
\]  
(9)

on the second subdomain. We use (9) to eliminate the unknown traction \( \gamma \) from (8) and vice versa. Solving the interface equations in (9) for \( \gamma \) yields

\[ \langle \gamma, N_{2,i} \rangle_{\sigma} = (f, N_{2,i})_{\Omega_2} - (\sigma(\hat{u}_2), \varepsilon(N_{2,i}))_{\Omega_2} - (\hat{u}_{2,\sigma}, N_{2,i})_{\Omega_2} & \forall i \in V(\sigma_2^h). \]  
(10)

Equation (10) defines a finite element approximation \( \gamma_2(\hat{u}_{2,\sigma}, u_2) \in S_{2,\sigma}^h \) of the interface traction in terms of \( \hat{u}_{2,\sigma} \) and \( u_2 \). It can be interpreted as variational recovery [2] of \( \gamma \) from a finite element solution. Then we approximate \( \gamma \) in (8) by the interpolant \( \Pi_1 \gamma_2 \in S_{1,\sigma}^h \). This yields the following system of equations on the first subdomain:

\[
\begin{align*}
(\hat{u}_{1,\sigma}, N_{1,i})_{\Omega_1} + (\sigma(\hat{u}_1), \varepsilon(N_{1,i}))_{\Omega_1} &= (f, N_{1,i})_{\Omega_1} + \langle \Pi_1 \gamma_2(\hat{u}_{2,\sigma}, u_2), N_{1,i} \rangle_{\sigma} & \forall i \in V(\sigma_1^h) \\
(\hat{u}_{1,0}, N_{1,i})_{\Omega_1} + (\sigma(\hat{u}_1), \varepsilon(N_{1,i}))_{\Omega_1} &= (f, N_{1,i})_{\Omega_1} & \forall i \in V(\Omega_1^h) \\
u_{1,\sigma} &= \Pi_1(u) u_{2,\sigma}
\end{align*}
\]  
(11)
Conversely, using (8) to eliminate \( \gamma \) from (9) we obtain an analogous equation on \( \Omega_2 \):

\[
\begin{cases}
(\ddot{u}_{2,\sigma}, N_{2,i})_{\Omega_2} + (\sigma(\mathbf{u}_2), \epsilon(N_{2,i}))_{\Omega_2} = (f, N_{2,i})_{\Omega_2} - \langle \Pi_2 \gamma_1 (\ddot{u}_{1,\sigma}, \mathbf{u}_1), N_{2,i} \rangle_{\Omega_2} & \forall i \in V(\sigma_h^2) \\
(\ddot{u}_{2,0}, N_{2,i})_{\Omega_2} + (\sigma(\mathbf{u}_2), \epsilon(N_{2,i}))_{\Omega_2} = (f, N_{2,i})_{\Omega_2} & \forall i \in V(\hat{\Omega}_2^h) \\
\mathbf{u}_{2,\sigma} = \Pi_2(\mathbf{u}) u_{1,\sigma}
\end{cases}
\]  

(12)

Let \( \mathbf{F}_i = (\mathbf{F}_{1,i}, \mathbf{F}_{1,0}) \) be the vector with elements

\[
\mathbf{F}_i^k = (f, N_{i,k})_{\Omega_i} - (\sigma(\mathbf{u}_i), \epsilon(N_{i,k}))_{\Omega_i} \quad \forall k \in V(\Omega_i^h). 
\]  

(13)

Then, the interface equation in (11) can be written as

\[
M_{1,\sigma} \ddot{u}_{1,\sigma} = \mathbf{F}_{1,\sigma} + M_{1,\sigma} P_1 \gamma_2 (\ddot{u}_{2,\sigma}, \mathbf{u}_2), 
\]  

(14)

whereas the matrix form of equation (10), which defines \( \gamma_2 \), is given by

\[
\mathbf{M}_{2,\sigma} \gamma_2 = \mathbf{F}_{2,\sigma} - M_{2,\sigma} \ddot{u}_{2,\sigma}. 
\]

Solving the latter for \( \gamma_2 \) yields

\[
\gamma_2 (\ddot{u}_{2,\sigma}, \mathbf{u}_2) = \mathbf{M}_{2,\sigma}^{-1} \mathbf{F}_{2,\sigma} - \mathbf{M}_{2,\sigma}^{-1} M_{2,\sigma} \ddot{u}_{2,\sigma}. 
\]

The algebraic form of (11) follows by substituting this result into (14):

\[
\begin{cases}
M_{1,\sigma} \ddot{u}_{1,\sigma} + M_{1,\sigma} P_1 M_{2,\sigma}^{-1} M_{2,\sigma} \ddot{u}_{2,\sigma} = \mathbf{F}_{1,\sigma} + M_{1,\sigma} P_1 \mathbf{M}_{2,\sigma}^{-1} \mathbf{F}_{2,\sigma} \\
M_{1,0} \ddot{u}_{1,0} = \mathbf{F}_{1,0} \\
\mathbf{u}_{1,\sigma} = P_1(\mathbf{u}) \mathbf{u}_{2,\sigma}
\end{cases}
\]  

(15)

Proceeding along the same lines we obtain an analogous algebraic form for (11):

\[
\begin{cases}
M_{2,\sigma} \ddot{u}_{2,\sigma} + M_{2,\sigma} P_2 M_{1,\sigma}^{-1} M_{1,\sigma} \ddot{u}_{1,\sigma} = \mathbf{F}_{2,\sigma} + M_{2,\sigma} P_2 \mathbf{M}_{1,\sigma}^{-1} \mathbf{F}_{1,\sigma} \\
M_{2,0} \ddot{u}_{2,0} = \mathbf{F}_{2,0} \\
\mathbf{u}_{2,\sigma} = P_2(\mathbf{u}) \mathbf{u}_{1,\sigma}
\end{cases}
\]  

(16)

4.4 Elimination of displacement continuity equations

Equations (15)–(16) remain coupled through their dependence on interface states from both subdomains. Under some additional assumptions on the matrix structure \( P_i(\mathbf{u}_i) \) can be effectively approximated by the interface interpolant \( P_i \) in which case the weak continuity equations in (15)–(16) are replaced by

\[
\mathbf{u}_{1,\sigma} = P_1 \mathbf{u}_{2,\sigma} \quad \text{and} \quad \mathbf{u}_{2,\sigma} = P_2 \mathbf{u}_{1,\sigma}, 
\]  

(17)
respectively. The key factor that enables such an approximation is to work with diagonal mass matrices. Thus, from now on we assume that (i) assembly is performed using node-based quadrature rules, which result in
\[ M_{i,\sigma} = \text{diag}(m^k_{i,\sigma}) \quad \text{and} \quad \mathcal{M}_{i,\sigma} = \text{diag}(\mathcal{m}^k_{i,\sigma}); \quad i = 1, 2, \]
and (ii) displacement continuity conditions are given by (17). For clarity we explain elimination of interface states in a two-dimensional setting. In this case matrix forms of interface transfer operators \( \Pi_i \) assume a particularly simple form with at most two non-zero elements per row. We explain the structure of \( P_1 \). Let \( x_{1,i} \in \sigma^h_1 \) be an arbitrary vertex on the interface of \( \Omega_1 \) and \( K_{2,k} \in \sigma^h_2 \) be the element from the interface of \( \Omega_2 \), which contains\(^2 \) \( x_{1,i} \). Since \( \sigma \) is one-dimensional, element \( K_{2,k} \) is an interval with endpoints \( x_{2,k_i-1} \) and \( x_{2,k_i} \), respectively. As a result, (1) reduces to the following sum
\[ \sum_{k \in V(K_{2,k_i})} \bar{u}^k_{2,\sigma} N_{2,k}(x_{1,i}) = \bar{u}^{k_{i-1}}_{2,\sigma} N_{2,k_i-1}(x_{1,i}) + \bar{u}^{k_i}_{2,\sigma} N_{2,k_i}(x_{1,i}) \tag{18} \]
Since basis functions form a partition of unity on every element, \( N_{2,k_i-1}(x_{1,i}) + N_{2,k_i}(x_{1,i}) = 1 \) and so, there exists \( 0 \leq \alpha_i \leq 1 \) such that \( N_{2,k_i-1}(x_{1,i}) = \alpha_{1,i} \) and \( N_{2,k_i}(x_{1,i}) = 1 - \alpha_{1,i} \). It follows that the matrix \( P_1 \) is given by
\[ (P_1)_{ij} = \begin{cases} \alpha_{1,i} & \text{if } j = k_i - 1 \\ 1 - \alpha_{1,i} & \text{if } j = k_i \\ 0 & \text{otherwise} \end{cases} \tag{19} \]
where \( K_{2,k_i} = [x_{2,k_i-1}, x_{2,k_i}] \) is the element from the interface on \( \Omega_2 \) containing vertex \( x_{1,i} \) from the interface on \( \Omega_1 \). Repeating the same arguments for \( \Pi_2 \) shows that
\[ (P_2)_{ij} = \begin{cases} \alpha_{2,i} & \text{if } j = k_i - 1 \\ 1 - \alpha_{2,i} & \text{if } j = k_i \\ 0 & \text{otherwise} \end{cases} \tag{20} \]
where \( K_{1,k_i} = [x_{1,k_i-1}, x_{1,k_i}] \) is the element from the interface on \( \Omega_1 \) containing vertex \( x_{2,i} \) from the interface on \( \Omega_2 \) and \( \alpha_{2,i} = N_{1,k_i-1}(x_{2,i}) \) and \( 1 - \alpha_{2,i} = N_{1,k_i}(x_{2,i}) \).

Since interior equations are fully decoupled from the interface equations we focus solely on the structure of the latter. Their right hand sides are given by
\[
(\bar{F}_{1,\sigma} + \mathcal{M}_{1,\sigma} P_1\mathcal{M}^{-1}_{2,\sigma} \bar{F}_{2,\sigma})_j = F^j_{1,\sigma} + \mathcal{m}^j_{1,\sigma} \left[ \alpha_{1,j} \frac{F^{k_{j-1}}_{2,\sigma}}{\mathcal{m}^{k_{j-1}}_{2,\sigma}} + (1 - \alpha_{1,j}) \frac{F^{k_j}_{2,\sigma}}{\mathcal{m}^{k_j}_{2,\sigma}} \right]
\]
\(^2\)If \( x_{1,i} \) is also a vertex in \( \sigma^h_1 \), then it is shared by two elements in \( \sigma^h_2 \). In this case we can take \( K_{2,k_i} \) to be either one of these two elements.
that is, for the interface equation on $\Omega_1$ and

$$
\left( \bar{F}_{2,\sigma} + M_{2,\sigma} P_2 M_{1,\sigma}^{-1} \bar{F}_{1,\sigma} \right)_j = F_{2,\sigma}^j + \bar{m}_{2,\sigma}^j \left[ \alpha_{2,j} \frac{F_{1,\sigma}^{k_j-1}}{m_{1,\sigma}^{k_j-1}} + (1 - \alpha_{2,j}) \frac{F_{1,\sigma}^{k_j}}{m_{1,\sigma}^{k_j}} \right],
$$

for the interface equation on $\Omega_2$.

Consider the terms involving displacements from the opposite sides of the interface, that is, $M_{1,\sigma} P_1 M_{2,\sigma} M_{2,\sigma} \bar{u}_{2,\sigma}$ in (15) and $M_{2,\sigma} P_2 M_{1,\sigma} M_{1,\sigma} \bar{u}_{1,\sigma}$ in (16). We have that

$$
\left( M_{1,\sigma} P_{1} M_{2,\sigma} M_{2,\sigma} \bar{u}_{2,\sigma} \right)_j = \bar{m}_{1,\sigma}^j \left[ \alpha_{1,j} \frac{m_{2,\sigma}^{k_j-1}}{m_{1,\sigma}^{k_j-1}} \bar{u}_{2,\sigma}^{k_j-1} + (1 - \alpha_{1,j}) \frac{m_{2,\sigma}^{k_j}}{m_{1,\sigma}^{k_j}} \bar{u}_{2,\sigma}^{k_j} \right],
$$

and

$$
\left( M_{2,\sigma} P_{2} M_{1,\sigma} M_{1,\sigma} \bar{u}_{1,\sigma} \right)_j = \bar{m}_{2,\sigma}^j \left[ \alpha_{2,j} \frac{m_{1,\sigma}^{k_j-1}}{m_{2,\sigma}^{k_j-1}} \bar{u}_{1,\sigma}^{k_j-1} + (1 - \alpha_{2,j}) \frac{m_{1,\sigma}^{k_j}}{m_{2,\sigma}^{k_j}} \bar{u}_{1,\sigma}^{k_j} \right].
$$

For shape-regular grids it is not unreasonable to expect that (see Fig.1)

$$
\frac{m_{2,\sigma}^{k_j-1}}{m_{2,\sigma}^{k_j}} \approx \frac{m_{2,\sigma}^{k_j-1}}{m_{1,\sigma}^{k_j}} := \mu_{2,\sigma}^j \quad \text{and} \quad \frac{m_{1,\sigma}^{k_j-1}}{m_{2,\sigma}^{k_j}} \approx \frac{m_{1,\sigma}^{k_j-1}}{m_{2,\sigma}^{k_j}} := \mu_{1,\sigma}^j.
$$

This allows us to exchange the order of interpolation and matrix multiplication in (21):

$$
\left( M_{1,\sigma} P_{1} M_{2,\sigma} M_{2,\sigma} \bar{u}_{2,\sigma} \right)_j = \bar{m}_{1,\sigma}^j \mu_{2,\sigma}^j \left[ \alpha_{1,j} \bar{u}_{2,\sigma}^{k_j-1} + (1 - \alpha_{1,j}) \bar{u}_{2,\sigma}^{k_j} \right] = \bar{m}_{1,\sigma}^j \mu_{2,\sigma}^j (P_1 \bar{u}_{2,\sigma}).
$$

Likewise, exchanging the order of operators in (22) gives

$$
\left( M_{2,\sigma} P_{2} M_{1,\sigma} M_{1,\sigma} \bar{u}_{1,\sigma} \right)_j = \bar{m}_{2,\sigma}^j \mu_{1,\sigma}^j \left[ \alpha_{2,j} \bar{u}_{1,\sigma}^{k_j-1} + (1 - \alpha_{2,j}) \bar{u}_{1,\sigma}^{k_j} \right] = \bar{m}_{2,\sigma}^j \mu_{1,\sigma}^j (P_2 \bar{u}_{1,\sigma}).
$$

From (17) it follows that $\bar{u}_{2,\sigma} = (P_1 \bar{u}_{2,\sigma})_j$ and $\bar{u}_{1,\sigma} = (P_2 \bar{u}_{1,\sigma})_j$. Using these identities we can eliminate $\bar{u}_{2,\sigma}$ from (21) and $\bar{u}_{1,\sigma}$ from (22) to obtain

$$
\left( M_{1,\sigma} P_{1} M_{2,\sigma} M_{2,\sigma} \right) \bar{u}_{2,\sigma} \approx (M_{1,\sigma} \mu_{2,\sigma}) \bar{u}_{1,\sigma}
$$

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure1.png}
\caption{Assumption (23) holds provided $(A_1 + A_2)/(A_2 + A_3) \approx (E_1 + E_2)/(E_2 + E_3)$ and $(B_4 + B_5)/B_5 \approx (E_4 + E_5)/E_5$.}
\end{figure}
and
\[
\left( M_{2,\sigma} P_2 \bar{M}^{-1}_{1,\sigma} M_{1,\sigma} \bar{u}_{1,\sigma} \right) \approx \left( M_{2,\sigma} \mu_{1,\sigma} \right) \ddot{u}_{2,\sigma}
\]
respectively. This decouples (15)–(16) into an independent equation
\[
\begin{aligned}
& \left( M_{1,\sigma} + \bar{M}_{1,\sigma} \mu_{2,\sigma} \right) \ddot{u}_{1,\sigma} = \bar{F}_{1,\sigma} + \bar{M}_{1,\sigma} P_1 \bar{M}_{2,\sigma}^{-1} \bar{F}_{2,\sigma} \\
& M_{1,0} \ddot{u}_{1,0} = \bar{F}_{1,0}
\end{aligned}
\] (24)
on \Omega_1, and another independent subdomain equation
\[
\begin{aligned}
& \left( M_{2,\sigma} + \bar{M}_{2,\sigma} \mu_{1,\sigma} \right) \ddot{u}_{2,\sigma} = \bar{F}_{2,\sigma} + \bar{M}_{2,\sigma} P_2 \bar{M}_{1,\sigma}^{-1} \bar{F}_{1,\sigma} \\
& M_{2,0} \ddot{u}_{2,0} = \bar{F}_{2,0}
\end{aligned}
\] (25)
on \Omega_2. The interface equations in each subdomain have the following component form:
\[
\begin{aligned}
& \left( m_{1,\sigma}^j + \bar{m}_{1,\sigma}^j \mu_{2,\sigma} \right) \ddot{u}_{1,\sigma}^j = F_{1,\sigma}^j + \bar{m}_{1,\sigma}^j \left[ \alpha_{1,j} \frac{F_{2,\sigma}^{k_j-1}}{m_{2,\sigma}^j} + \left( 1 - \alpha_{1,j} \right) \frac{F_{2,\sigma}^{k_j}}{m_{2,\sigma}^j} \right]; \ j \in V(\sigma_1^h)
\end{aligned}
\] (26)
and
\[
\begin{aligned}
& \left( m_{2,\sigma}^j + \bar{m}_{2,\sigma}^j \mu_{1,\sigma} \right) \ddot{u}_{2,\sigma}^j = F_{2,\sigma}^j + \bar{m}_{2,\sigma}^j \left[ \alpha_{2,j} \frac{F_{1,\sigma}^{k_j-1}}{m_{1,\sigma}^j} + \left( 1 - \alpha_{2,j} \right) \frac{F_{1,\sigma}^{k_j}}{m_{1,\sigma}^j} \right]; \ j \in V(\sigma_2^h).
\end{aligned}
\] (27)

Modification of subdomain mass matrices in (26)–(27) can be interpreted as their completion to bulk mass matrices on \( S_1^{h,\sigma} \cup S_2^{h,\sigma} \).

4.5 Fully discrete partitioned equations

We discretize (26)–(27) in time using second central difference
\[
\ddot{u}_i(t, x) \approx \frac{u_i(t + \Delta t, x) - 2u_i(t, x) + u_i(t - \Delta t, x)}{\Delta t^2}.
\]
Let \( u_i^{n+1} \in S_i^{h,\sigma} \), \( u_i^n \in S_i^{h,\sigma} \) and \( u_i^{n-1} \in S_i^{h,\sigma} \) denote finite element approximations of \( u_i \) at \( t_n + \Delta t \), \( t_n \), and \( t_{n-1} = t_n - \Delta t \), respectively, \( \bar{F}_i^{n+1}(u_i) = (u_i^{n+1} - 2u_i^n + u_i^{n-1})/\Delta t^2 \), and \( (\bar{F}_i)^n \) be the force vector (13) evaluated at \( u_i^n \). Then, for given \( u_i^n \) and \( u_i^{n-1} \), the fully discrete partitioned formulation is to find \( u_i^{n+1} \) such that
\[
\begin{aligned}
& \left( M_{1,\sigma} + \bar{M}_{1,\sigma} \mu_{2,\sigma} \right) \bar{D}^{n+1}(u_{1,\sigma}) = (\bar{F}_{1,\sigma})^n + \bar{M}_{1,\sigma} P_1 \bar{M}_{2,\sigma}^{-1} (\bar{F}_{2,\sigma})^n \\
& M_{1,0} \bar{D}^{n+1}(u_{1,0}) = (\bar{F}_{1,0})^n
\end{aligned}
\] (28)
and \( u_2^{n+1} \) such that
\[
\begin{aligned}
& \left( M_{2,\sigma} + \bar{M}_{2,\sigma} \mu_{1,\sigma} \right) \bar{D}^{n+1}(u_{2,\sigma}) = (\bar{F}_{2,\sigma})^n + \bar{M}_{2,\sigma} P_2 \bar{M}_{1,\sigma}^{-1} (\bar{F}_{1,\sigma})^n \\
& M_{2,0} \bar{D}^{n+1}(u_{2,0}) = (\bar{F}_{2,0})^n
\end{aligned}
\] (29)
for the finite element approximations \( u_i^{n+1}, i = 1, 2 \) of the subdomain solutions at \( t_{n+1} \).
5 Equivalence to a monolithic discretization on matching interface grids

If Ω₁^h and Ω₂^h are such that interface grids match then Ω₁∪Ω₂ is a conforming partition of Ω and S^h = S₁^h ∪ S₂^h is a conforming finite element subspace of H¹(Ω). The corresponding monolithic formulation of (2) is

\[ M \ddot{D}^{n+1}v = (\ddot{F})^n. \]

where M and \( \ddot{F} \) are a diagonal mass matrix and force vector assembled using \( S^h \). Partitioning of mesh nodes into interface and subdomain nodes induces partitioning of the solution vector \( v \) into coefficient vectors \( v_\sigma, v_{1,0} \) and \( v_{2,0} \) corresponding to interface and interior subdomain nodes, respectively. As a result, we can write the monolithic problem in the following block diagonal form:

\[
\begin{cases}
M_\sigma \ddot{D}^{n+1}(v_\sigma) = (\ddot{F}_\sigma)^n \\
M_{1,0} \ddot{D}^{n+1}(v_{1,0}) = (\ddot{F}_{1,0})^n \\
M_{2,0} \ddot{D}^{n+1}(v_{2,0}) = (\ddot{F}_{2,0})^n
\end{cases}
\] (30)

**Theorem 1** Assume that interface grids \( \sigma_1^h \) and \( \sigma_2^h \) are matching and interface displacements at all previous time steps coincide:

\[
(\vec{u}_{1,\sigma})' = (\vec{u}_{2,\sigma})' \quad \nu = 1, 2, \ldots, n - 1, n.
\] (31)

Then the partitioned solution \((\vec{u}_{1,\sigma}, \vec{u}_{1,0})^{n+1}, (\vec{u}_{2,\sigma}, \vec{u}_{2,0})^{n+1}\) coincides with the solution \( \vec{v}^{n+1} = (\vec{v}_\sigma, \vec{v}_{1,0}, \vec{v}_{2,0})^{n+1} \) of the monolithic problem: \( \vec{v}_\sigma^{n+1} = \vec{u}_{1,\sigma}^{n+1} = \vec{u}_{2,\sigma}^{n+1}, \vec{v}_{1,0}^{n+1} = \vec{v}_{2,0}^{n+1} \) and \( \vec{v}_{2,0}^{n+1} = \vec{v}_{2,0}^{n+1} \).

Proof. For clarity we present the proof for the two-dimensional formulation (24)–(25) and skip the time step index. Owing to the assumption that interface grids on \( \Omega_1 \) and \( \Omega_2 \) match, it follows that the area mass matrices \( \overline{M}_{1,\sigma} \) and \( \overline{M}_{2,\sigma} \) are identical, i.e., they have the same dimension and with proper renumbering of their elements we can write

\[
\overline{m}_{1,\sigma} = \overline{m}_{2,\sigma} \quad \forall j \in V(\sigma_1^h) \equiv V(\sigma_2^h)
\]

For matching interface grids we also have that \( P_1 = P_2 = I \). As a result, the interface equations assume the form

\[
(m_{1,\sigma}^j + m_{2,\sigma}^j) u_{1,\sigma}^j = F_{1,\sigma}^j + F_{2,\sigma}^j \quad \forall j \in V(\sigma_1^h)
\]

and

\[
(m_{2,\sigma}^j + m_{1,\sigma}^j) u_{2,\sigma}^j = F_{2,\sigma}^j + F_{1,\sigma}^j \quad \forall j \in V(\sigma_2^h),
\]

respectively. Thus, for matching interface grids (28)–(29) has the form

\[
\begin{cases}
(M_{1,\sigma} + M_{2,\sigma}) \vec{u}_{1,\sigma} = \vec{F}_{1,\sigma} + \vec{F}_{2,\sigma} \\
M_{1,0} \vec{u}_{1,0} = \vec{F}_{1,0}
\end{cases} \\
\begin{cases}
(M_{2,\sigma} + M_{1,\sigma}) \vec{u}_{2,\sigma} = \vec{F}_{2,\sigma} + \vec{F}_{1,\sigma} \\
M_{2,0} \vec{u}_{2,0} = \vec{F}_{2,0}
\end{cases}
\] (32)
It immediately follows that $\vec{u}_{1,0} = \vec{v}_{1,0}$ and $\vec{u}_{2,0} = \vec{v}_{2,0}$. On the other hand, it is easy to see that for matching interface partitions, the monolithic volume interface mass matrix is sum of the volume interface mass matrices on $\Omega_1$ and $\Omega_2$, i.e., $M_\sigma = M_{1,\sigma} + M_{2,\sigma}$. Furthermore, if (31) holds, a direct calculation shows that the monolithic interface force vector is sum of the interface force vectors on $\Omega_1$ and $\Omega_2$: $\vec{F}_\sigma = \vec{F}_{1,\sigma} + \vec{F}_{2,\sigma}$. Therefore, $\vec{u}_{1,\sigma}$ and $\vec{u}_{2,\sigma}$ solve an identical equation, which coincides with the monolithic interface equation and so, $\vec{u}_{1,\sigma} = \vec{u}_{2,\sigma} = \vec{v}_\sigma$. □

6 Convergence rates

We use the manufactured solution

$$u = \left(\sin(5\pi x) \cos(3\pi y) \log(1 + t); 4x^4 \cos(4\pi y) \sqrt{t + 2}\right)^T$$ (33)

to estimate numerical convergence rates of the algorithm. We assume linear homogeneous isotropic solid with $\mu = 0.01$, $\lambda = 0.02$ dyne/cm$^2$ and density 1 g/cm$^3$. Substitution of (33) into the governing equations yields the problem data. The domain $\Omega = [0,1]^2$ is divided into two subdomains using a vertical and a slanted interface; see Fig. 2. Each subdomain is meshed independently and the interface grids are non matching.

<table>
<thead>
<tr>
<th>Mesh 1</th>
<th>Mesh 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>12 × 20</td>
<td>20 × 20</td>
</tr>
<tr>
<td>24 × 40</td>
<td>40 × 40</td>
</tr>
<tr>
<td>48 × 80</td>
<td>80 × 80</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Error/Rate</th>
<th>$|u - u^*<em>1|</em>{0,\Omega_1}$</th>
<th>$|u - u^*<em>1|</em>{0,\Omega_2}$</th>
<th>$|u - u^*<em>2|</em>{1,\Omega_1}$</th>
<th>$|u - u^*<em>2|</em>{1,\Omega_2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.97e-03/1.95</td>
<td>5.12e-04/2.01</td>
<td>6.42e-03/2.01</td>
<td>1.59e-03/2.01</td>
<td>1.30e-01/1.00</td>
</tr>
<tr>
<td>2.06e-03/1.95</td>
<td>2.58e-02/1.00</td>
<td>2.59e-01/1.11</td>
<td>5.29e-01/1.00</td>
<td>1.30e-01/1.00</td>
</tr>
<tr>
<td>6.42e-03/2.01</td>
<td>1.59e-03/2.01</td>
<td>1.59e-03/2.01</td>
<td>1.30e-01/1.00</td>
<td>1.30e-01/1.00</td>
</tr>
<tr>
<td>5.12e-04/2.01</td>
<td>1.30e-01/1.00</td>
<td>5.29e-01/1.00</td>
<td>1.30e-01/1.00</td>
<td>1.30e-01/1.00</td>
</tr>
<tr>
<td>2.58e-02/1.00</td>
<td>1.59e-03/2.01</td>
<td>1.59e-03/2.01</td>
<td>1.30e-01/1.00</td>
<td>1.30e-01/1.00</td>
</tr>
</tbody>
</table>

Table 1: Errors and convergence rates using a vertical interface and uniform meshes at $t = 0.25$ s.
Table 2: Errors and convergence rates using a slanted interface and non uniform meshes at $t = 0.25 s$.

We observe in Tables 1 and 2 that by using a coincidental interface with nonmatching vertices and temporal step sizes on the order of $h$, the rate of convergence is second order regardless of the interface orientation.

6.1 Equivalence to a monolithic solution for matching interface grids

To confirm numerically Theorem 1 we use the same interface configurations as before, but consider grids with matching interface nodes. In this case the union $\Omega_1^h \cup \Omega_2^h$ defines a conforming mesh partition of $\Omega$. The difference in solutions, shown in Table 3, left, are equivalent up to roundoff whether computed through the monolithic formulation or using the algorithm based on variational flux recovery.

6.2 Preservation of linear displacements

The patch test [7] requires a method to recover a certain class of solutions. This section verifies that our method is capable of reproducing linear displacement fields exactly. We consider $\Omega = [-1, 1]^2$ with a vertical and a slanted interface and non matching interface grids; see Fig.3. The exact solution is

$$u = (-5x + 50y, 33x - 22y)$$

(34)

We assume linear homogenous isotropic solid with $\mu = 1.5$, $\lambda = 7$ dyn/cm$^2$ and density $1 \text{ g/cm}^3$. As before, substitution of (34) into the governing equations yields the problem data. Table 3, right, confirms that the new algorithm recovers the linear displacement field up to machine precision, i.e., it passes a patch test for non matching interface grids.
7 CONCLUSIONS

We have presented a new explicit method for elastodynamic problems with interfaces, which enables partitioned solution of the equations. Numerical studies show that the method passes a linear patch test and is second order accurate. If the interface grids have matching nodes then the method recovers a solution of the monolithic discretization.

REFERENCES


LEVEL SET ANALYSIS OF TOPOLOGY OPTIMIZATION IN 2D STRUCTURES USING BOUNDARY ELEMENT METHOD

PAULO CEZAR VITORIO JUNIOR*, EDSON DENNER LEONEL†

* PhD Student at School of Engineering of São Carlos – University of São Paulo
  Av. Trabalhador Sãocarlense, 400 – São Carlos - SP / Brazil
  paulovitorio@usp.br

† Professor at School of Engineering of São Carlos – University of São Paulo
  Av. Trabalhador Sãocarlense, 400 – São Carlos - SP / Brazil
  edleonel@sc.usp.br

Key words: Topology optimization. Level set method. Boundary element method.

Abstract. The present study applies an evolutionary approach to determine the optimal geometry of 2D structures by coupling the Boundary Element Method (BEM) and Level Set Method (LSM). Topology optimization is a scientific domain that aims to design structural components with appropriate safety level using the minimum amount of material. Therefore, this type of optimization aims to achieve the optimal structural geometry, i.e. the shape that leads to the minimum requirement of material respecting constraints related to stress state at any point of the structure.

As this problem involves moving boundary conditions, BEM becomes well adapted to solve him due to its non-requirement of domain mesh. Moreover, the velocity of moving boundaries is determined by Von Mises stress. Thus, as BEM is efficient to evaluate internal stresses and displacements, BEM is recommended to model this complex and important engineering problem.

The model developed in this study consists of three parts: mechanical model, topology optimization scheme and structure reconstruction. The mechanical model is composed by BEM algebraic equations. Singular and hyper-singular integral representations are available. The topology optimization is performed using LSM. The body’s geometry is determined by the Level Set (LS) function evaluated at its zero level. The reconstruction process concerns the remeshing procedure. As the boundary moves at each iteration, the body’s geometry change and a new mesh has to be constructed.

During the optimization process, the developed model introduces automatically internal cavities which are included according to the intensity of Von Mises stress. The zero level set contours describe internal and external geometries which are discretized by linear boundary elements.

The developed numerical model was applied to several benchmarks available in literature. According to the obtained results good agreement was observed among them demonstrating the efficiency of the developed model.
1 INTRODUCTION

The structural design tries to couple conditions that are normally opposite: safety and economy. In most part of structural designs, the optimal solution, i.e., the solution that couples perfectly these conditions, is not easily available as it depends on the structural dimensions, material strengths and structural system adopted. In order to solve consistently this optimization problem some scientific domains have recently emerged. Among them, it is worth to mention the topology optimization which is a scientific domain that aims to design the geometry of structural components with appropriate safety level using the minimum amount of material. This type of optimization aims to achieve the optimal structural geometry, i.e. the shape that leads to the minimum requirement of material respecting constraints related to stress state at any structural point.

The mechanical modelling can be performed by analytical methods based on the knowledge of the theory of elasticity. However, the solutions provided by the theory of elasticity are limited to simplified geometries and boundary conditions. To consider complex geometries, boundary conditions and material constitutive relations numerical methods are required. The present study applies the Boundary Element Method (BEM), which is recognized as a robust and efficient numerical technique capable to handle accurately several types of mechanical problems. Due to its mesh dimension reduction and accuracy in determining stress concentration fields, this numerical method becomes well adapted to solve topology optimization problems.

Topology optimization was initially proposed by [1] which analysed the material distribution into a fixed domain. The optimal structural geometry was determined by defining a material density, which may vary from void (no material) until full presence of material. An intermediate material density may also be assumed. [2] introduced the bubble method, which is based on the insertion of new holes in the structure and the subsequent use of a shape optimization method to determine their optimal size and shape. The simple evolutionary method (ESO) presented by [3] progressively removes material from low stress regions based on some rejection criteria. This technique was widely applied coupled to finite element models.

An efficient approach in the topology optimization domain consists of the Level Set Method (LSM). This method was initially proposed to model the movement of curves [4,5]. In the structural context, the desired curve is the body’s boundary and its movement or evolution characterizes the new structural geometry. The LSM is based on the solution of Hamilton-Jacobi differential equations and it represents the body’s geometry and its evolution by the zero level of an important function called Level Set (LS) [4-9]. A topology optimization approach based on coupling BEM-LSM was proposed by [10]. During the optimization process, the proposed procedure introduces internal cavities automatically, in order to accelerate the numerical analysis. The zero level set describes the internal and external structural geometry. The initial boundaries as well as its evolutions are represented by NURBS, which smooth the BEM mesh. The internal cavities are punctually included where the von Mises stress is lower than an established threshold.

In the present study, the topology optimization analysis is developed by the coupling BEM-LSM. The proposed algorithm consists of three parts: mechanical model, topology optimization procedure and structure reconstruction. The mechanical model is composed by BEM algebraic equations in which singular and hyper-singular integral representations can be applied. The topology optimization procedure is performed using LSM. The body’s geometry is described by the LS function evaluated at its zero level. The velocities of moving boundaries are determined according to the intensity of Von
Mises stress at each boundary node. Finally, the reconstruction process concerns the remeshing procedure. As the boundary moves at each iteration, the body’s geometry change and a new mesh has to be constructed. One application is presented in this study in order to illustrate the potential of application of the numerical tool developed.

2 BOUNDARY ELEMENT METHOD

The BEM has been widely applied in engineering fields such as contact mechanics and fracture mechanics due to its high accuracy and robustness in modelling strong stress concentration. In addition to this inherent ability of BEM, its mesh dimension reduction makes this numerical method well adapted to solve topology optimization problems. The integral equations required by BEM can be obtained from the equilibrium equation. Considering a two-dimensional homogeneous elastic domain, Ω, with boundary, Γ, the equilibrium equation, written in terms of displacements, is given by:

\[ u_{i,j} + \frac{1}{1-2\nu} u_{j,i} + \frac{b_i}{\mu} = 0 \]  

(1)

where \( \mu \) is the shear modulus, \( \nu \) is the Poisson’s ratio, \( u_i \) are components of the displacement field, and \( b_i \) are body forces.

Using Betti’s theorem, the singular integral representation, written in terms of displacements, is obtained (without body forces) as follows:

\[ c_{ij}(s,f)u_j(s) + \int_\Gamma P_{ij}^*(s,f)u_j(f)\,d\Gamma = \int_\Gamma U_{ij}^*(s,f)p_j(f)\,d\Gamma \]  

(2)

where \( p_j \) and \( u_j \) are tractions and displacements at the boundary, respectively, the free term \( c_{ij} \) is equal to \( \delta_{ij}/2 \) for smooth boundaries and \( P_{ij}^* \) and \( U_{ij}^* \) are the fundamental solutions for tractions and displacements written for the source point \( s \) [11-13].

Differentiating the Eq. (2) with respect to the directions \( x, y \) an integral equation written in terms of strains is obtained. Applying, for instance, the generalized Hooke’s law the integral representation of stresses for a boundary source point is obtained. Then, the Cauchy formula can be applied to obtain the hyper-singular integral equation which is written in terms of tractions as follows:

\[ \frac{1}{2} p_j(s) + \eta_k \int_\Gamma S_{kij}^*(s,f)u_k(f)\,d\Gamma = \eta_k \int_\Gamma D_{kij}^*(s,f)p_k(f)\,d\Gamma \]  

(3)

\( S_{kij}^* \) and \( D_{kij}^* \) contain the new kernels computed from \( P_{ij}^* \) and \( U_{ij}^* \) respectively [11-13] and \( \eta_k \) is the outward normal vector.

To assemble the system of BEM equations, as usual, Eq. (2) or (3) are transformed into algebraic relations by discretizing the boundary into elements along which displacements and tractions are approximated. In the present study, only linear discontinuous boundary elements are applied. After determining the displacement and tractions fields at the boundary, internal values for displacements, stresses and strains can be achieved. Internal displacements are determined using the integral Eq. (2) with the source point \( s \) located at the interior of the domains. In this case, the free term \( c_{ij} \) becomes \( \delta_{ij} \). On the other hand, the stress field at internal nodes is obtained through the integral representation of stresses [11-13].
3 LEVEL SET METHOD

The LSM is a robust technique to simulate and determine the movement of curves in different physical scenarios. This method represents a particular curve (or surface) $\Gamma$ and its evolution along time as the zero level (zero level set) of a function $\phi$, which is called Level Set function. Figure 1 illustrates a classical example of one LS function. This figure presents the evolution of a circle, whose level set function is a cylinder.

![Level Set Function](image)

The LS equation is represented in the following form.

$$\phi_t + \vec{V} \cdot \nabla \phi = 0$$

(4)

In which $\phi_t$ is the partial time derivative, $\vec{V}$ is the velocity field on the grid points and $\nabla \phi$ is the gradient of the $\phi$ function which represents the partial derivatives according to $x$ and $y$ directions.

By solving Eq.(4), using finite differences, for instance, the values of $\phi$ are determined for all points that belong to the analysed domain. Therefore, based on $\phi$ values, the new structural boundaries are achieved. The LSM is capable to eliminate complexities of movement of curves such as singularities, weak solutions, shocks formation, non-stable conditions of entropy and topology changes that involve interfaces. In the context of numerical analysis, the LSM uses natural and accurate computational procedures. The method represents accurately corners and geometry ends as well as topological discontinuities.

4 TOPOLOGY OPTIMIZATION BY COUPLING BEM-LSM

As previously mentioned, in the present study the topology optimization analysis is developed by the coupling BEM-LSM. The proposed algorithm consists of three parts: mechanical model, topology optimization procedure and structure reconstruction.

Considering the mechanical model, the BEM requires the structure discretization by nodes and elements in order to determine stresses and displacements at the structural boundaries. The structural boundaries are described by the level set function evaluated at its zero level which is discretized by linear discontinuous boundary elements.

In order to solve the topology optimization problem, an Eulerian mesh defined by a rectangular grid is required. The structure is immersed into this grid and its geometrical coordinates are part of Eulerian mesh [14]. For each grid point the LS function is determined. This function is defined as the orientated distance between the grid points and the structural boundary. The grid points positioned internally to the boundary are
represented by negative values of \( \phi(x) \). Positive values of \( \phi(x) \) indicate points positioned outside the boundary. Null value of \( \phi(x) \) defines the zero level set and, consequently, the structural boundary. Thus, the function \( \phi(x) \) is written as presented in Eq. (5), where \( \partial \Omega = \Gamma \).

\[
\begin{align*}
0 < \phi(x) & \quad \text{to } \forall x \in \Omega \setminus \partial \Omega \\
\phi(x) = 0 & \quad \text{to } \forall x \in \partial \Omega \\
\phi(x) < 0 & \quad \text{to } \forall x \in D \setminus \Omega 
\end{align*}
\] (5)

In this study, Eq.(4) is solved using upwind differences method. As \( \phi_t \) and \( \phi \) are known, a velocity function \( V \) must be defined. The velocity function depends on the Von Mises stress, Eq.(6), which is evaluated for each grid node.

\[
\sigma_{vm} = \frac{1}{\sqrt{2}} \sqrt{(\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_1 - \sigma_3)^2}
\] (6)

in which \( \sigma_1, \sigma_2 \) and \( \sigma_3 \) are the principal stresses.

Based on the Von Mises stress evaluated by BEM, inefficient materials are progressively removed as well as materials are added where necessary. The regions that satisfy the Eq. (7) have removal material condition.

\[
\sigma_{vm} < RR \sigma_{\text{max}}
\] (7)

where \( RR \) is the material reduction module defined in the beginning of analysis and \( \sigma_{\text{max}} \) is the maximum von Mises stress calculated considering the initial structure. Similarly, the regions where material must be added are identified considering Eq. (8). If Eq.(8) is true material is added.

\[
\sigma_{vm} > \min(\sigma_{\text{max}}, \sigma_y)
\] (8)

in which \( \sigma_y \) represents the yield stress of the material in question.

The circular cavities are included into the structural domain in points where the von Mises stress reaches minimum values. Just one cavity per iteration is inserted.

The LS equation is discretized through upwind difference method and the velocities for each grid point are calculated according to the intensity of Von Mises stress. The upwind difference method calculates the partial derivatives of the LS function in relation to \( x \) and \( y \) directions. The time interval used by the method is considered as fictitious. Another input parameter is the number of evolutions carried out at each iteration.

In order to define the velocity range according to the intensity of Von Mises stress, stress intervals are defined. The intervals are characterized as a function of \( \sigma_{vm}, \, RR, \, \sigma_y \) and \( \sigma_{\text{max}} \). The velocity criteria adopted in this study is presented in Eq. (9).

\[
\sigma_{vm} \in [0, \sigma_{t1}]: \sigma_{t1} = 0.5RR \sigma_{\text{max}}, \quad VN = -1
\]
\[
\sigma_{vm} \in [\sigma_{t1}, \sigma_{t2}]: \sigma_{t2} = 0.9RR \sigma_{\text{max}}, \quad VN \in [-1; 0]
\]
\[
\sigma_{vm} \in [\sigma_{t2}, \sigma_{t3}]: \sigma_{t3} = 0.95 \min(\sigma_{\text{max}}, \sigma_y), \quad VN = 0
\]
\[
\sigma_{vm} \in [\sigma_{t3}, \sigma_{t4}]: \sigma_{t4} = \min(\sigma_{\text{max}}, \sigma_y), \quad VN \in [0; 1]
\]
\[
\sigma_{vm} \in [\sigma_{t4}, \infty]: \ VN = 1
\] (9)

The negative velocity indicates that the grid node movement is orientated for inside the boundary. Therefore, negative velocities eliminate inefficient material. Similarly, the movement of grid nodes orientated to the outside boundary direction occurs when
the velocity calculated is positive. In this condition, material is added in the structure. Null velocity indicates non-movement of grid/boundary nodes. The criteria presented in Eq.(9) is graphically represented in Figure 2.

The LSM requires an update mesh procedure, i.e. the mesh considered in the previous iteration must be eliminated and a new mesh is constructed at the new zero level set. Thus, a remeshing procedure must be performed. The developed algorithm is capable to recognize the amount of curves present in the structure (outer and internal boundaries) and achieve the zero level as final boundary. Then, the zero level set is discretized in nodes and linear boundary elements for further analysis of the BEM. This procedure is performed by linear interpolation of the $\phi$ function along the entire grid. This process is repeated until the desired optimal area be reached. In order to help computational implementation by the reader, a flowchart is introduced in Figure 3.
The developed topology optimization scheme was applied to a cantilever beam subjected to a vertical load. Figure 4 illustrates the initial structure including dimensions (in metre) and load position and intensity.
The following material properties were adopted: 210 GPa, 0.3 and 280 MPa, they are, respectively, Young’s modulus ($E$), Poisson ratio ($\nu$), and yield stress $\sigma_y$. 

Figure 4: Cantilever beam

Figure 5: Evolution process
The reduction factor adopted is $RR = 0.25$ and 5 evolutions for each step were performed by the upwind difference method. The fictitious time used is $\Delta t = 0.003$. The grid is composed by a regular mesh of $67 \times 45$ points, i.e. 3015 points. The values of $\Delta x$ and $\Delta y$ are equal to 0.1 m. The goal of this application is the determination of the structural geometry that contains 40% of the initial area.

Figure 5 shows the structural geometry evolution from the initial step until de convergence. The creation of internal cavities must be observed which accelerated the convergence. It is worth to mention that the obtained results agree with classical references.

Figure 6 illustrates the reduction of area during the iterative process. According to this figure a smooth convergence is observed which indicates the numerical stability of the developed procedure.

Figure 7 presents the evolution of the level set function until the convergence. According to this figure the geometry complexity of the level set function must be mentioned.

![Figure 6: Reduction of area during the numerical procedure](image_url)
6 CONCLUSIONS

In this study the analysis of topology optimization was performed by the coupling BEM-LSM. In such type of coupling the advantages of each method are used to obtain optimal structural geometries in an efficient and accurate manner.

According to the results obtained good agreement was observed between the results achieved by the developed BEM-LSM coupled model and the responses available in literature. Smooth and stable numerical convergence was also observed indicating that the numerical model is stable.

However, some improvements should be performed into the developed model. Among these improvements, the parallelization of the code is due in course in order to improve the computational time consuming. In addition to that, the application of new velocity functions is recommended.

ACKNOWLEDGEMENTS

Sponsorship of this research project by the São Paulo Research Foundation (FAPESP), grant number 2014/18928-2 and National Council of Technological and Scientific Development (CNPq), is greatly appreciated.
REFERENCES


A NEW MULTISCALE SIMULATION OF WATER FLOWS IN HETEROGENEOUS POROUS MEDIA

A.S. FRANCISCO¹, L.M. SACRAMENTO² AND J.J. TRUJILLO³

¹Universidade Federal Fluminense
Avenida dos Trabalhadores, 420
27255-125, Volta Redonda-RJ, Brazil
afrancisco@metal.eeimv.frrr.ufv.br

² Universidade Federal Fluminense
Avenida dos Trabalhadores, 420
27255-125, Volta Redonda-RJ, Brazil
leonardo_sacra@hotmail.com

³ Universidade Federal Fluminense
Avenida dos Trabalhadores, 420
27255-125, Volta Redonda-RJ, Brazil
jjuarezt@hotmail.com

Key words: Multiscale method, Finite difference, Porous media.

Abstract. We apply the multiscale finite difference method (MsFDM) to approximate pressures and velocities in highly heterogeneous porous media flows. Heterogeneities in the permeability coefficient are assumed to occur on a wide range of length scales. The multiscale method allows incorporate fine-scale information on the permeability into coarse-scale simulations. In this direction, local, multiscale basis functions are defined for representing the discrete solutions in subdomains. Similar procedure was employed successfully in using hybridized mixed finite elements for the spatial discretization of the equations. Here, global and local flow problems are approximated using the finite difference method through coarse and fine grids, respectively. Such hierarchical scheme is expected to be an inexpensive alternative to solve the global flow problem on fine grids only. Numerical results are presented to illustrate the computational performance of the multiscale method.

1 INTRODUCTION

Fluid flow problems are adequately modeled combining the mass conservation equation with the Darcy law, broadly used to model porous media flows. The Darcy law states that the volumetric flux of a fluid in porous media is proportional to the pressure gradient. It defines permeability of the porous media as a property that quantifies the capacity of the medium to displace fluids. The Darcy law in addition to the mass conservation derives then a second order elliptic equation.
The development of multiscale numerical procedures for second order elliptic equations related to porous media flow problems has attracted the attention of several research groups (Ganis and Yotov, 2009; Kippe et al, 2008). Multiscale procedures can provide large computational savings when applied as direct solvers for problems governed by physical processes occurring on a wide range of heterogeneity scales.

Francisco et al (2011) have developed a hierarchical two-scale approach in which new simulation strategies are incorporated. The hierarchical approach is applied to solve iteratively global coarse and local fine problems. In Francisco et al (2014) we find the MsFDM method based on an overlapping iterative domain decomposition that uses a finite difference discretization for approximating the heterogeneous porous media flow problem.

In this paper, new numerical experiments using the MsFDM method are presented in order to assess the results for heterogeneous permeability fields considering several grid refinement problems.

This paper is organized as follows. In the second section is presented the governing equations of the problem; a short description of the approximating technique for solving the problem is found in the third section; in the fourth, comparisons of numerical results are presented; and conclusions appear in the last section.

2 GOVERNING EQUATIONS

For simplicity in presentation, it is considered a bounded domain $\Omega \subset \mathbb{R}^2$, with Lipschitz boundary $\partial \Omega$. The global problem for one-phase fluid flows in porous media [8] is given

$$\nabla \cdot u = f,$$  
(1)

$$u = -K(x)\nabla p,$$  
(2)

where $K$ is the intrinsic permeability tensor, $u$ is the Darcy velocity, $p$ is the pressure of the fluid, and $f$ is the source/sink term. Typical boundary conditions that occur in porous medium flow problems are Dirichlet and Neumann, which are respectively written as

$$p = p_b \text{ on } \Gamma_D \text{ and } u \cdot n = u_b \text{ on } \Gamma_N,$$  
(3)

where $\partial \Omega = \Gamma_D \cup \Gamma_N$, $\Gamma_D \cap \Gamma_N = \emptyset$ and $n$ is an outward unit vector normal to $\partial \Omega$. The global problem is solved in the level of a fine grid.

3 THE MULTISCALE FINITE DIFFERENCE METHOD

A natural way to solve second order elliptic equations in parallel is to divide the domain over which the equations are redefined into subdomains, and to solve local problems in parallel (Dawson et al., 1991). The major difficulties with such procedure involve imposing interface values between subdomains (Min and Yang, 2007). We overcome these difficulties using overlapping subdomains.
The spatial domain $\Omega$ is divided into partitions $\{\Omega_j\}, \; j = 1: M$, associated to the coarse grid. A hierarchical two-scale configuration is posed in the domain. In each subdomain $\Omega_j$, the local problem is given by

$$\nabla \cdot (-K(x) \nabla p_j) = f, \quad (4)$$

where $p_j$ is the local solution for the pressure. The discrete solution for the local problem (4) is given by

$$p_j = \sum_{i=1}^{4N} A_{ji} \psi_{ji}, \quad (6)$$

where the coefficients $A_{ji}$ are the pressures on the neighboring edge points and $\psi_{ji}$, $i = 1,...,4N$ be basic functions associated with the subdomain, where $N$ denotes the number of nodal points on each edge of the subdomain. The basic functions are approximations of the following equation:

$$\nabla \cdot (-K(x) \nabla \psi_{ji}) = f, \quad (5)$$

with boundary conditions given by $\psi_{bi} = 1$ on the $i$-edge point, and $\psi_{bi} = 0$ on all other edge points.

To perform the iterative domain decomposition procedure, basic functions have to be pre-computed. In the discretization of the Eq. (5) is employed the finite difference method, while the preconditioned conjugate gradient is used for solving the resulting algebraic equation system (Burden and Faires, 2004).

4 NUMERICAL EXPERIMENTS

We perform experiments with the slab geometry problem, which are common in simulation of single-phase flow problems. We consider a physical domain size of $128m \times 128m$, the source/sink term $f = 0$, and the number of nodal points on each edge of the subdomain $N = 1$.

The experiments use a two-dimensional homogeneous and heterogeneous permeability fields, with average permeability of $1.0 \times 10^{-11} m^3/s/kg$. The left and right boundary conditions are of Dirichlet type ($p_b = 1.0$ Pa and 0.0 Pa, respectively), and the top and bottom boundary conditions are of Neumann type ($u_b = 0.0$ m/s). The slab geometry is depicted in the Fig. 1.

3.1 Homogeneous problem

For the homogeneous slab problem, we consider a computational fine grid of $128 \times 128$. 
We perform a comparison of the MsFDM solution with the analytical solution, considering several coarse grids.

![Figure 1 – The slab geometry.](Image)

Table 1 shows results of number of iterations and global error according to the coarse grid refinement. As the coarse grid is refined, the number of iterations increases and the global error decreases.

<table>
<thead>
<tr>
<th>Coarse grid</th>
<th>Global error</th>
<th>Iteration</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 x 8</td>
<td>1.70E-02</td>
<td>401</td>
</tr>
<tr>
<td>16 x 16</td>
<td>8.07E-03</td>
<td>377</td>
</tr>
<tr>
<td>32 x 32</td>
<td>5.23E-03</td>
<td>286</td>
</tr>
<tr>
<td>64 x 64</td>
<td>3.47E-03</td>
<td>51</td>
</tr>
</tbody>
</table>

Oscillations in the resulting pressure field are provided by the numerical resolution using the MsFDM method. These numerical errors are observed on the interface of subdomains due to the coarse approximation using the number of nodal points on each edge of the subdomain $N = 1$. However, as the coarse grid is refined, we can note that the numerical errors present a tendency to decrease. The Fig. 2 illustrates the resulting pressure fields according to the coarse grid refinement.

### 3.2 Heterogeneous problem

For the heterogeneous slab problem, we consider a computational fine grid of $128 \times 128$ and the heterogeneous permeability field given by $K(x) = K_0 \exp(\delta \varepsilon(x))$, where $\varepsilon(x)$ is a self-similar field (Gaussian), characterized on Borges et al. (2008), $K_0$ is the average permeability and $\delta$ is chosen in order to set a highly permeability ratio.

We perform a comparison of the MsFDM solution with the solution obtained by using a mixed finite element combined with a iterative domain decomposition method (Douglas et al,
1997). Global error and number of iterations are evaluated with different permeability ratios, considering the coarse grid of 64 x 64.

![Figure 2](image-url)  

Figure 2 – The resulting pressure field for coarse grid of: (a) 8 x 8; (b) 16 x 16; (c) 32 x 32; and (d) 64 x 64.

In Table 2 is shown the results of this experiment. As the permeability ratio increases, the global error and number of iterations present larger values.

<table>
<thead>
<tr>
<th>$K_{\text{max}}/K_{\text{min}}$</th>
<th>Global error</th>
<th>Iteration</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^1$</td>
<td>2.44E-02</td>
<td>953</td>
</tr>
<tr>
<td>$10^3$</td>
<td>6.04E-02</td>
<td>1139</td>
</tr>
<tr>
<td>$10^6$</td>
<td>1.29E-01</td>
<td>1391</td>
</tr>
</tbody>
</table>

Table 2 – Results for the permeability ratios.
5 CONCLUSIONS

- The results for the heterogeneous case do not present a satisfactory order of accuracy for global errors. We believe that MsFDM will present better results after implementing post-processing strategies.
- We conclude that the MsFDM is promising when considering the implementation of parallel programming.

REFERENCES


REALISTIC CFD SIMULATION OF COMPRESSIBLE FLOW INSIDE THE CORE OF THE PEBBLE BED REACTOR HTR-10

Leorlen Y. Rojas Mazaira 1,3, Dany S. Dominguez 2, Carlos R. G. Hernández 1, Abel Gámez Rodríguez 1, Carlos Alberto Brayner de Oliveira Lira 3

1 Instituto Superior de Tecnologías y Ciencias Aplicadas, Av. Salvador Allende esq. Luaces – CP 10600. La Habana, Cuba
leored1984@gmail.com, cgh@instec.cu, agamezgmf@gmail.com

2 Pós-graduação em Modelagem Computacional, Universidade Estadual de Santa Cruz, Rodovia Jorge Amado, Km 16, Salobrinho, CEP 45662-900, Ilhéus, BA, Brasil
dsdominguez@gmail.com

3 Universidade Federal de Pernambuco, Av. Luiz Freire 1000, Cid. Universitaria, CEP 50740-540, Recife, PE, Brasil
cabol@ufpe.br

Key words: HTR-10, Pebbled Bed Reactor, FCC.

Abstract. High-temperature gas-cooled reactors (HTGRs) have the potential to be used as possible energy generation sources in the near future, owing to their inherently safe performance by using a large amount of graphite, low power density design, and high conversion efficiency. However, safety is the most important issue for its commercialization in nuclear energy industry. It is very important for safety design and operation of an HTGR to investigate its thermal–hydraulic characteristics. In this article, it was performed the thermal–hydraulic simulation of compressible flow inside the core of the pebble bed reactor HTR-10 (High Temperature Reactor) using Computational Fluid Dynamics (CFD). The realistic approach was used, where every closely packed pebble is realistically modelled. Due to the high computational cost is impossible simulate the full core; therefore, the geometry used is a FCC (Face Centered Cubic) cell with the height of the core, with a total of 82 pebbles. The input data used were taken from the thermal–hydraulic IAEA Benchmark [1]. The results show the profiles of velocity and temperature of the coolant in the core, and the temperature distribution inside the pebbles. The maximum temperatures in the pebbles do not exceed the allowable limit for this type of nuclear fuel.

1 INTRODUCTION

The pebble bed nuclear reactor is a high temperature gas cooled reactor design, and one of the main candidates for next generation nuclear power plants. In these reactors, the fuel is contained within graphite spheres, which form a randomly packed bed in the graphite-walled core region. Key features are the passive safety of the reactor, higher thermal efficiency due to higher coolant outlet temperatures, and the possibility
of on-line refueling by extracting pebbles at the bottom of the bed and adding pebbles to the top.

The HTR-10 was selected as reference of HTGR experimental reactor by OIEA [1] and it is selected for the present simulations. This reactor was designed by the Institute of Nuclear and New Energy Technology of Tsinghua University in Mainland China. It is a graphite-moderated and helium-cooled reactor with a pebble bed core of spherical fuel elements. The schematic for the HTR-10 is shown in Fig. 1 [1]. The design characteristics are listed in Table 1. In this reactor, the coolant gas would flow around the randomly distributed spheres at the core.

![Figure 1 Schematic representation of HTR-10.](image)

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thermal power, MW</td>
<td>10</td>
</tr>
<tr>
<td>Primary helium pressure, MPa</td>
<td>3.0</td>
</tr>
<tr>
<td>Average helium outlet temperature, K</td>
<td>973</td>
</tr>
<tr>
<td>Average helium inlet temperature, K</td>
<td>523</td>
</tr>
<tr>
<td>Helium mass flow rate, kg/s</td>
<td>4.32</td>
</tr>
<tr>
<td>Power, MW</td>
<td>3.0</td>
</tr>
</tbody>
</table>

Table 1. Main design and operational parameters of the HTR-10
The pebble bed core with a diameter of 180 cm and an average height of 197 cm is composed of spherical fuel elements (equilibrium core) or mixture of fuel balls and graphite dummy balls without fuel (initial core). Each fuel ball contains an inner fuel zone of 50 mm diameter, with the coated TRISO (tristructural isotropic) particles, and a fuel-free graphite shell of 5 mm thickness. In addition, each graphite ball has the same diameter (60 mm) as the fuel ball. The bottom structure of the reactor core is designed to be conical to facilitate free movement of the balls by gravity. The whole active core zone is enclosed by graphite neutron reflectors, which are categorized as the top, side, and bottom reflectors.

There are two main approaches for the CFD simulation of the close packed pebble geometry: the porous approach and the realistic approach. In the porous approach, an averaged concept of porosity is applied to simulate the close packed geometry.

Recently, with the dramatic progress in computational capability, several researchers have adopted a realistic approach for the pebble geometry to simulate the distribution characteristics of gas flow within close packed pebble geometries. In the realistic approach, every pebble is modelled [2, 3, 4, 5, 6, 7] this model with higher simulation detail allows to study the non-uniform thermal-hydraulic behavior within the closely packed pebble geometry, in order to locate hot spots and calculate the maximum temperatures in the fuel elements more accurately.

In Pebbled Bed Reactor (PBR), the pebbles are randomly arranged in the core and the porosity varies radially and axially, mostly in the near wall region. Due to the geometric complexity of a random structure, a simplification used very frequently is the structure of regular arrangements: Simple Cubic (SC), Body Centered Cubic (BCC) and Face Centered Cubic (FCC).

In one of the first realistic CFD simulation of a PBR [3], the flow through pebbles was simulated in a BCC lattice as a segment of a PBR core with the large eddy simulation (LES) technique. In [4], advantages and disadvantages of realistic and porous approaches in the close packed pebbles with the BCC arrangement are investigated, comparing CFD simulation results.

The effects of two different packing arrangements (BCC and FCC) on the thermal–hydraulic characteristics within a PBR core were investigated in [2]. These characteristics include flow acceleration in the pebble gap, vortices and flow separation in the pores, anisotropic variation of Nusselt number and temperature on a pebble surface, among others. Due to more compact lattice, the flow acceleration in the FCC-arrangement pebbles is stronger than that in the BCC arrangement, resulting in the higher heat transfer capability and subsequently the lower pebble temperature.

In a previous work [8], it was performed the simulation of whole core of HTR-10 using the porous approach. The temperature at outlet obtained with the porous model was used for a realistic model with a FCC cell with three layers (in the most critical region). That methodology overestimate the temperature reach by the fuel. In this work is performed a simulation with a FCC cell with the total height of the HTR-10, to improve the model, and obtain the axial temperature distribution of the core of a PBR.

At the next section is described the geometric and numeric model. In section 3 are presented the results and in section 4, some conclusions and suggestions to future
works are given.

2 MODEL DESCRIPTIONS

2.1 Computational domain and boundary conditions

The simulation domain consists in a FCC arrangement with 2 mm spacing between pebbles, with 187 cm of height. The fuel was assumed as a homogeneous sphere. The geometry presented has 41 layers of pebbles with 21 full pebbles, 80 half pebbles and 84 quarter pebbles, with a total of 82 pebbles. The pebbles are stationary in the core. The geometry is shown in the Figure 2.

Helium gas is set to be injected from the top of pebbles and an extra flow length is added to the last layer of pebbles in order to conserve the fully developed flow condition at the outlet. It was applied a symmetry boundaries on the four side surfaces. Mass flow input for this simulation is calculated proportionally to the total coolant’s mass flow in the core (4.32 kg/s) and it is 0.01344 kg/s. Pebbles emit a uniform volumetric heat flux, depending on the position in the core. From the power
distribution of the initial core of HTR-10 shown in Table 2 [1]. The power distribution used for this simulation is just the first column of Table 2 that corresponds to the centerline of the core.

**Table 2. Power distribution of the initial core of HTR-10 (W/cm³).**

<table>
<thead>
<tr>
<th>Z\R(cm)</th>
<th>6.5</th>
<th>13</th>
<th>19</th>
<th>25</th>
<th>39</th>
<th>50</th>
<th>60</th>
<th>70.5</th>
<th>83.5</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>1.79</td>
<td>1.79</td>
<td>1.79</td>
<td>1.78</td>
<td>1.75</td>
<td>1.71</td>
<td>1.68</td>
<td>1.65</td>
<td>1.63</td>
<td>1.65</td>
</tr>
<tr>
<td>36</td>
<td>2.12</td>
<td>2.11</td>
<td>2.10</td>
<td>2.08</td>
<td>2.06</td>
<td>2.01</td>
<td>2.46</td>
<td>1.89</td>
<td>1.86</td>
<td>1.89</td>
</tr>
<tr>
<td>54</td>
<td>2.46</td>
<td>2.45</td>
<td>2.44</td>
<td>2.41</td>
<td>2.35</td>
<td>2.26</td>
<td>2.18</td>
<td>2.10</td>
<td>2.07</td>
<td>2.09</td>
</tr>
<tr>
<td>72</td>
<td>2.61</td>
<td>2.60</td>
<td>2.58</td>
<td>2.56</td>
<td>2.49</td>
<td>2.43</td>
<td>2.33</td>
<td>2.27</td>
<td>2.22</td>
<td>2.22</td>
</tr>
<tr>
<td>90</td>
<td>2.78</td>
<td>2.77</td>
<td>2.75</td>
<td>2.72</td>
<td>2.65</td>
<td>2.56</td>
<td>2.46</td>
<td>2.39</td>
<td>2.30</td>
<td>2.30</td>
</tr>
<tr>
<td>108</td>
<td>2.84</td>
<td>2.84</td>
<td>2.82</td>
<td>2.79</td>
<td>2.71</td>
<td>2.55</td>
<td>2.45</td>
<td>2.33</td>
<td>2.26</td>
<td>2.26</td>
</tr>
<tr>
<td>126</td>
<td>2.66</td>
<td>2.65</td>
<td>2.63</td>
<td>2.60</td>
<td>2.56</td>
<td>2.49</td>
<td>2.38</td>
<td>2.29</td>
<td>2.22</td>
<td>2.21</td>
</tr>
<tr>
<td>144</td>
<td>2.49</td>
<td>2.49</td>
<td>2.47</td>
<td>2.44</td>
<td>2.37</td>
<td>2.29</td>
<td>2.19</td>
<td>2.08</td>
<td>2.01</td>
<td>2.00</td>
</tr>
<tr>
<td>162</td>
<td>2.36</td>
<td>2.35</td>
<td>2.33</td>
<td>2.30</td>
<td>2.22</td>
<td>2.11</td>
<td>2.02</td>
<td>1.93</td>
<td>1.85</td>
<td>1.82</td>
</tr>
<tr>
<td>180</td>
<td>2.33</td>
<td>2.32</td>
<td>2.29</td>
<td>2.25</td>
<td>2.16</td>
<td>2.02</td>
<td>1.93</td>
<td>1.83</td>
<td>1.73</td>
<td>1.66</td>
</tr>
<tr>
<td>187</td>
<td>1.53</td>
<td>1.53</td>
<td>1.51</td>
<td>1.48</td>
<td>1.42</td>
<td>1.33</td>
<td>1.27</td>
<td>1.21</td>
<td>1.13</td>
<td>0</td>
</tr>
</tbody>
</table>

Properties of helium surrounding the pebbles were calculated according to the German Safety Guide. Where d is the pebble’s diameter, \( k_{Helium} \) is helium thermal conductivity and the dynamic viscosity \( \eta_{Helium} \), which can be calculated as [9]:

\[
k_{Helium} = 2.682 \cdot 10^{-3} \left(1 + 1.123 \cdot 10^{-3} \cdot P\right) \cdot T^{0.7} \left(10^{-7} \cdot P\right)
\]  

\[
\eta_{Helium} = 3.674 \cdot 10^{-7} \cdot T^{0.7}
\]

Where \( P \) is the pressure and \( T \) the temperature, and \( C_{pHelium} = 5195 \) J/kg K are the heat capacity. The density of fuel was \( \rho_{fuel} = 1720 \) kg/m³, the conductivity and specific heat of fuel was assumed a constant average as, \( k_{fuel} = 24.7 \) W/m·K and \( C_{p fuel} = 1690 \) J/kg K.

### 2.2 Numerical methods

For the simulations of the flow field and convective heat transfer ANSYS CFX 14 [10] was used. To resolve the phenomena in three dimensions, the Reynolds-averaged equations for compressible flows were solved. For modeled turbulence the \( k - \omega \) SST model is used due to its enhanced wall treatment and the benefits of combining \( k - \varepsilon \) in the free stream and using \( k - \omega \) in the near wall region. In packed beds can be observed a variation of length scales and Reynolds numbers. In the advection term of the RANS simulation, the High Resolution scheme was used, in which the value of \( \beta \) is computed locally to be as close to 1 as possible without violating the boundedness principles. The High Resolution scheme is therefore both accurate and bounded. For the calculation of the turbulence quantities, the first-order upwind scheme was used.
Unstructured grids are used in the whole geometry due to the complexity. Mesh independent calculations are also performed to verify that the simulation results presented below are mesh independent. Two final cases with 36447378 and 51995823 of elements were compared. The maximum deviations in the maximum temperature are less than 1.0% for the cases with different mesh sizes.

3 RESULTS AND DISCUSSION

Regarding safety, temperature of the fuel elements during normal operation and transients is a very important topic that should be studied in a nuclear reactor to assure the integrity of materials composing the fuel element and avoid the release of fission products. Leakage of gaseous fission products is produced through the fuel element by means of diffusion processes; when temperatures raise the diffusion increases [11, 12]. During accidents, air could enter inside the core due to depressurization of the reactor and cause integrity loss of TRISO particles because of gasification of the graphite matrix [12, 13]. In this case, if temperature is very high the integrity of TRISO particles can be more affected due to manufacturing defects, internal pressure of fission gases, etc. [12, 14]. The limit temperature at which the TRISO particles begin releasing the gaseous fission products is 1873 K (1600 °C) [11, 12, 13, 14, 15].

Figure 3 gives the axial temperature contours on the center plane of the core including the pebbles and helium. At right side of Figure 3 a zoomed view of the last pebbles with the maximum temperatures with a local scale of temperature. The calculations show that the core temperature increases along the axial direction achieving a maximum value of 1116 K (843 °C) in the fuel and an average temperature at outlet of 1048 K (775 °C).

Due to the combination of the axial power distribution and the helium flow direction, the maximum core temperatures obtained by various participants in [1] are in the range of 800-1000 °C, at the bottom of the cylindrical part of the pebble bed.

The results of this work are consistent with this benchmark. The maximum temperature of the fuel is in the range of other simulations, and the temperature reached by helium in the centerline is higher than the average outlet temperature achieved in the reactor. However, it was to be expected that in the realistic 3D CFD model proposed, the maximum temperatures are near or superiors that obtained by the 2D models developed by others researchers [1]. Some assumptions made in this model, such as, a homogeneous sphere and constant properties for fuel, may underestimate the temperatures reached.

In Figure 4 is shown the first and last section of the core with the surface temperature of the pebbles and the streamlines of the flow. It shows that the gas flows onto the front of the pebble, causing the higher heat transfer and the lower surface temperature. On the rear of a pebble, the flow detachment would result in the inferior heat transfer and simultaneously the higher surface temperature, as clearly shown. The higher temperature difference (~60 K) between the front and the rear locations for the pebble at the outlet respect to the inlet (~50 K) is revealed. This predicted result implies the higher anisotropic characteristic in the heat transfer at
pebble surface in different position at axial direction in the core. The variation in density due to the variation of temperature of the helium results in higher flow acceleration and consequently higher turbulent mixing affecting the pebble.

Figura 4. Helium and fuel temperatures at centerline plane.

Figura 5. Helium streamlines and fuel surface temperatures at inlet and outlet regions.
4 CONCLUSIONS

The realistic CFD simulation of compressible flow of an FCC cell with the height of the core of the pebble bed reactor HTR-10 was carried out. Some important conclusions could be drawn from the present simulations and would be presented as follows:

- The steady-state temperature distribution of the HTR-10 at the axial centerline of the core were obtained, and the maximum temperatures reached by the fuel 1116 K (843 °C) are much lower than the limit value for TRISO fuel 1873 K (1600 °C).

- The complexity of the flow results in changes in the temperature difference at pebble surface that can be a design problem in reactors larger than the HTR-10.

Nevertheless, this realistic CFD model of the whole height of HTR-10 has many advantages, some assumptions made has to be improve. In future works will be use a model of the fuel with the graphite shell and the properties dependent of temperature. Another improvement will be consider the loss of coolant.

ACKNOWLEDGEMENTS

This work has been carried out in the framework of technological international cooperation supported the MES-CAPES 114/11 project titled “Study of Advanced Nuclear Systems pebble bed type with TRISO fuel for the transmutation of long-lived waste and hydrogen production with high temperature technologies”. Also, the authors thank to CAPES (Higher Education Co-ordination Agency of Education Ministry of Brazil) and to the Brazilian National Institute of Science and Technology of Innovative Nuclear Reactors.

REFERENCES


NEUMANN SOLUTIONS TO FRACTIONAL LAMÉ-CLAPEYRON-STEFAN PROBLEMS WITH HEAT FLUX OR CONVECTIVE BOUNDARY CONDITIONS

DOMINGO A. TARZIA

¹CONICET – Depto. Matemática, Univ. Austral
Paraguay 1950, S2000FZF Rosario, Argentina
E-mail: DTarzia@austral.edu.ar

Key words: Fractional Lamé-Clapeyron-Stefan Problem, Phase-change Problem, Neumann Solution, Wright Function, Mainardi Function, Explicit Solution.

Abstract. In this paper, generalized Neumann solutions for the two-phase fractional Lamé-Clapeyron-Stefan problems for a semi-infinite material are obtained with constant initial condition, and a boundary condition at the fixed face $x=0$ given by a heat flux or a convective (Robin) condition. In these problems, the two governing diffusion equations and a governing condition for the free boundary include a fractional time derivative in the Caputo sense of order $0<\alpha<1$. When $\alpha \to 1$ we recover the classical Neumann solutions for the two-phase Lamé-Clapeyron-Stefan problem through the error function, given in:

(i) Tarzia, Quart. Appl.Math., 39 (1981), 491-497, for a heat flux boundary condition at the fixed face $x=0$ when an inequality for the coefficient which characterizes the heat flux boundary condition is satisfied;
(ii) Tarzia, MAT – Serie A, 8 (2004), 21-27, for a convective boundary condition at the fixed face $x=0$ when an inequality for the coefficient which characterizes the convective boundary condition is satisfied.

1 INTRODUCTION

In the last decades the fractional differential equations were developed [10, 14, 16, 18-21] and in the recent years some works on the fractional Lamé-Clapeyron-Stefan problem were published [1, 8, 12, 13, 22-24, 32, 33].

In this paper, generalized Neumann solutions for the two-phase fractional Lamé-Clapeyron-Stefan problems for a semi-infinite material are obtained with constant initial temperature, and a boundary condition at the fixed face $x=0$ given by a heat flux or a convective (Robin) condition. Recently, a generalized Neumann solution for the two-phase fractional Lamé-Clapeyron-Stefan problem for a semi-infinite material with constant initial temperature, and a constant temperature condition at the fixed face $x=0$ was given in [24].

In these problems, the two governing diffusion equations and a governing condition for the free boundary include a fractional time derivative in the Caputo sense of order
0 < \alpha < 1 \text{ which is defined in [4]:}

\begin{equation}
D^\alpha f(t) = \frac{1}{\Gamma(1-\alpha)} \int_0^t \frac{f'(\tau)}{(t-\tau)^\alpha} \, d\tau \quad \text{for } 0 < \alpha < 1
\end{equation}

\begin{equation}
= f'(t) \quad \text{for } \alpha = 1
\end{equation}

where \( \Gamma \) is the Gamma function defined by:

\begin{equation}
\Gamma(x) = \int_0^\infty t^{x-1} \exp(-t) \, dt.
\end{equation}

Now, we define two functions (Wright and Mainardi functions) which are very important in order to obtain the explicit solutions in the following Sections.

The Wright function is defined in [34]:

\begin{equation}
W(z; \alpha, \beta) = \sum_{n=0}^{\infty} \frac{z^n}{\Gamma(n\alpha + \beta)}, \quad z \in \mathbb{C}, \quad \alpha > -1, \quad \beta \in \mathbb{R}.
\end{equation}

and the Mainardi function is defined in [10]:

\begin{equation}
M_\nu(z) = W(-z; -\nu, 1 - \nu) = \sum_{n=0}^{\infty} \frac{(-z)^n}{\Gamma(-n\nu + 1 - \nu)}, \quad z \in \mathbb{C}, \quad \nu < 1
\end{equation}

which is a particular case of the Wright function. Some basic properties are given by:

\begin{equation}
\frac{\partial W}{\partial z}(z; \alpha, \beta) = W(z; \alpha, \alpha + \beta)
\end{equation}

\begin{equation}
W(-x; -\frac{1}{2}, 1) = \text{erfc}\left(\frac{x}{2}\right), \quad 1 - W(-x; -\frac{1}{2}, 1) = \text{erf}\left(\frac{x}{2}\right)
\end{equation}

\begin{equation}
D^\alpha \left(t^\beta\right) = \frac{\Gamma(1+\beta)}{\Gamma(1+\beta-\alpha)} t^{\beta-\alpha}.
\end{equation}

Moreover, for the classical Lamé-Clapeyron-Stefan problem there exist thousands of papers on the subject, for example the first published papers [15, 26], the books [2,3,5-7, 9, 11, 17, 25, 31] and a large bibliography given in [28]. A review on explicit solutions with moving boundaries was given in [30].
In Section 2, we will obtain a generalized Neumann solution for the two-phase fractional Lamé-Clapeyron-Stefan problem for a semi-infinite material with constant initial condition, and a heat flux boundary condition at the fixed face \( x = 0 \). When \( \alpha \to 1^- \) we recover the Neumann solution for the classical two-phase Lamé-Clapeyron-Stefan problem through the error function, given in [27], when an inequality for the coefficient which characterizes the heat flux boundary condition is satisfied.

In Section 3, we will obtain a generalized Neumann solution for the two-phase fractional Lamé-Clapeyron-Stefan problem for a semi-infinite material with constant initial condition, and a convective (Robin) boundary condition at the fixed face \( x = 0 \). When \( \alpha \to 1^- \) we also recover the Neumann solution for the classical two-phase Lamé-Clapeyron-Stefan problem through the error function, given in [29], when an inequality for the coefficient which characterizes the convective boundary condition is satisfied.

2 THE TWO-PHASE FRACTIONAL LAMÉ-CLAPEYRON-STEFAN PROBLEM (MELTING PROCESS) WITH A HEAT FLUX BOUNDARY CONDITION AT THE FIXED FACE

We consider the following melting process:

**Problem** (FFP\(_\alpha\)) Find the free boundary \( x = s(t) \), and the temperature \( T = T(x,t) \) such that the following equations and conditions are satisfied \( (0 < \alpha < 1) \):

\[
\begin{align*}
D^\alpha T - \lambda^2_s T_{sx} &= 0, \quad x > s(t), \quad t > 0, \\
D^\alpha T_i - \lambda^2_i T_{ix} &= 0, \quad 0 < x < s(t), \quad t > 0, \\
s(0) &= 0, \\
T_s(x,0) &= T_s(+\infty,t) = T_f, \quad x > 0, \quad t > 0, \\
T_i(s(t),t) &= T_f, \quad t > 0, \\
k_s T_s(s(t),t) - k_i T_i(s(t),t) &= \rho \ell D^{\alpha} s(t), \quad t > 0, \\
k_i T_i(0,t) &= \frac{q_0}{t^{\alpha/2}}, \quad t > 0,
\end{align*}
\]

where \( \lambda^2_s = \frac{k_s}{\rho c_s}, \lambda^2_i = \frac{k_i}{\rho c_i} \).
Theorem 1 Let \( T_i < T_f \) be.

a) If the coefficient \( q_0 \) satisfies the inequality:

\[
q_0 > \frac{k_s(T_f - T_i)}{\lambda_i \Gamma(1-\alpha/2)},
\]

then there exists an instantaneous phase-change (melting) process and the problem \((\text{FFP}_{\alpha})\) has the generalized Neumann explicit solution given by:

\[
T_i(x,t) = T_f + \frac{q_0 \lambda_i \Gamma(1-\alpha/2)}{k_i} \left[ W\left(-\frac{x}{\lambda_i \Gamma^{\alpha/2}}; \frac{\alpha}{2}, 1\right) - W\left(-\lambda_{\xi_{fa}}\xi^{-\alpha/2}; \frac{\alpha}{2}, 1\right) \right]
\]

\[
T_s(x,t) = T_i + (T_f - T_i) \frac{W\left(-\frac{x}{\lambda_i \Gamma^{\alpha/2}}; \frac{\alpha}{2}, 1\right)}{W\left(-\lambda_{\xi_{fa}}\xi^{-\alpha/2}; \frac{\alpha}{2}, 1\right)},
\]

\[
s(t) = \xi_{fa} \lambda_i \Gamma^{\alpha/2},
\]

where the coefficient \( \xi = \xi_{fa} > 0 \) is the solution of the following equation:

\[
F_{\xi_{fa}}(x) = \frac{\Gamma\left(1+\frac{\alpha}{2}\right)}{\Gamma\left(1-\frac{\alpha}{2}\right)} x, \quad x > 0
\]

with

\[
F_{1a}(x) = \frac{q_0 \Gamma(1-\alpha/2)}{\rho \ell \lambda_s} M_{\alpha/2}(\lambda x) - \frac{k_s(T_f - T_i)}{\rho \ell \lambda_s^2} F_{2a}(x).
\]

\[
F_{2a}(x) = \frac{M_{\alpha/2}(x)}{W\left(-x; \frac{\alpha}{2}, 1\right)}.
\]

b) If the coefficient \( q_0 \) satisfies the inequalities

\[
0 < q_0 \leq \frac{k_s(T_f - T_i)}{\lambda_i \Gamma(1-\alpha/2)},
\]
then the problem \( (\text{FFP}_a) \) is a fractional diffusion problem for the initial solid phase whose solution is given by:

\[
T_s(x,t) = T_i + \frac{q_0 \lambda_1 (1-\alpha/2)}{k_s} W \left( -\frac{x}{\lambda_s t^{\alpha/2}}, -\frac{\alpha}{2}, 1 \right), \quad x > 0, \quad t > 0.
\]  

(24)

**Theorem 2** Let \( T_i < T_f \) be. If the coefficient \( q_0 \) satisfies the inequality (16) then the solution of the problem \( (\text{FFP}_a) \) converges to the solution of the classical Lamé-Clapeyron-Stefan problem \( (\text{FFP}_1) \) when \( \alpha \to 1^- \), and then we recover the classical Neumann explicit solution and the inequality for the coefficient which characterized the heat flux at \( x = 0 \) obtained for \( \alpha = 1 \) in [27], that is:

\[
q_0 > \frac{k_s(T_f - T_i)}{\sqrt{\pi\alpha s}}.
\]  

(25)

3 **THE TWO-PHASE FRACTIONAL LAMÉ-CLAPEYRON-STEFAN PROBLEM (SOLIDIFICATION PROCESS) WITH A CONVECTIVE BOUNDARY CONDITION AT THE FIXED FACE**

We consider the following solidification process:

**Problem \( (\text{FCP}_a) \)** Find the free boundary \( x = s(t) \), and the temperature \( T = T(x,t) \) such that the following equations and conditions are satisfied \((0 < \alpha < 1)\):

\[
D^\alpha T_i - \lambda_1^2 T_{i,x} = 0, \quad s(t) < x, \quad t > 0,
\]

(26)

\[
D^\alpha T_s - \lambda_2^2 T_{s,x} = 0, \quad 0 < x < s(t), \quad t > 0,
\]

(27)

\[
s(0) = 0,
\]

(28)

\[
T_s(x,0) = T_s(+\infty,t) = T_i > T_f, \quad x > 0, \quad t > 0,
\]

(29)

\[
T_s(s(t),t) = T_f, \quad t > 0,
\]

(30)

\[
T_i(s(t),t) = T_f, \quad t > 0,
\]

(31)

\[
k_s T_s(s(t),t) - k_i T_i(s(t),t) = \rho \epsilon D^\epsilon s(t), \quad t > 0,
\]

(32)

\[
k_s T_s(0,t) = \frac{h_0}{t^{\alpha/2}} (T_s(0,t) - T_x), \quad t > 0,
\]

(33)
where \( \lambda_s^2 = \frac{k_s}{\rho c_s} \), \( \lambda_i^2 = \frac{k_i}{\rho c_i} \).

**Theorem 3** Let \( T_S < T_f < T_i \) be.

a) If the coefficient \( h_0 \) satisfies the inequality:

\[
h_0 > \frac{k_i(T_i - T_f)}{\lambda_i(T_f - T_S) \Gamma(1 - \frac{\alpha}{2})}, \tag{34}
\]

then there exists an instantaneous phase-change (solidification) process and the problem (FCP\(_\alpha\)) has the generalized Neumann explicit solution given by:

\[
T_s(x, t) = T_f - (T_f - T_S) \left[ 1 - \frac{k_s}{h_0 \lambda_s \Gamma(1 - \frac{\alpha}{2})} + 1 - W\left( -\frac{x}{\lambda t^{\alpha/2}}; -\frac{\alpha}{2}, 1 \right) \right], \tag{35}
\]

\[
= T_S + (T_f - T_S) \left[ 1 - \frac{k_s}{h_0 \lambda_s \Gamma(1 - \frac{\alpha}{2})} + 1 - W\left( -\frac{x}{\lambda t^{\alpha/2}}; -\frac{\alpha}{2}, 1 \right) \right],
\]

\[
= T_f + (T_i - T_f) \left[ -W\left( -\frac{x}{\lambda t^{\alpha/2}}; -\frac{\alpha}{2}, 1 \right) \right] \tag{36}
\]

\[
= T_f + (T_i - T_f) \left[ -W\left( -\frac{x}{\lambda t^{\alpha/2}}; -\frac{\alpha}{2}, 1 \right) \right] \tag{36}
\]

\[
s(t) = \xi_{ca} \lambda_i t^{\alpha/2}, \tag{37}
\]

where the coefficient \( \xi = \xi_{ca} > 0 \) is the solution of the following equation:
\[ F_{c_x}(x) = \rho c \frac{\Gamma \left(1 + \frac{\alpha}{2}\right)}{\Gamma \left(1 - \frac{\alpha}{2}\right)} x, \quad x > 0 \]  

with

\[ F_{c_x}(x) = \frac{k_s(T_f - T_x)}{\lambda_s} F_{4a}(\frac{x^\alpha}{\lambda_s}) - \frac{k_i(T_i - T_f)}{\lambda_i} F_{2a}(x) \]  

where

\[ F_{4a}(x) = \frac{\mathcal{M}_{2\alpha}(x)}{h_0 \lambda_s \Gamma(1 - \frac{\alpha}{2}) + 1 - W \left( -x; -\frac{\alpha}{2}, 1 \right)} \]  

b) If the coefficient \( h_0 \) satisfies the inequalities

\[ 0 < h_0 \leq \frac{k_i(T_i - T_f)}{\lambda_i(T_f - T_x) \Gamma(1 - \frac{\alpha}{2})}, \]  

then the problem \((\text{FCP}_\alpha)\) is a fractional diffusion problem for the initial liquid phase whose solution is given by:

\[ T_i(x, t) = T_x + \frac{T_i - T_x}{k_i} \left[ \frac{k_i}{h_0 \lambda_s \Gamma(1 - \frac{\alpha}{2})} + 1 - W \left( -\frac{x}{\lambda_i t^{\alpha/2}}; -\frac{\alpha}{2}, 1 \right) \right], \quad x > 0, \quad t > 0 \]  

**Theorem 4** Let \( T_x < T_f < T_i \) be. If the coefficient \( h_0 \) satisfies the inequality (34) then the solution of the problem \((\text{FCP}_\alpha)\) converges to the classical solution of the problem \((\text{FCP})\) when \( \alpha \to 1 \). and then we recover the classical Neumann explicit solution and the inequality for the coefficient \( h_0 \) which characterized the convective (Robin) boundary condition at \( x = 0 \) obtained for \( \alpha = 1 \) in [29], that is:

\[ h_0 > \frac{k_i}{\sqrt{\pi \alpha_i}} \frac{T_i - T_f}{T_f - T_x}. \]
4 CONCLUSIONS
- We have obtained generalized Neumann solutions for two two-phase fractional Lamé-Clapeyron-Stefan problems for a semi-infinite material with constant initial condition, when a heat flux or a convective (Robin) boundary condition is imposed on the fixed face $x = 0$.
- The explicit solutions are given through the Wright and Mainardi functions.
- When $\alpha \to 1^-$, we recover the two classical Neumann solutions (which are equivalents among them) for the corresponding classical two-phase Lamé-Clapeyron-Stefan problem given through the error function, and also the inequalities for the corresponding coefficients which characterized the heat flux or the convective boundary condition at $x = 0$.

AKNOWLEDGEMENTS
The present work has been partially sponsored by the Project PIP No 0534 from CONICET-UA, Rosario, Argentina, and Grant AFORS FA9550-14-1-0122.

REFERENCES
[27] Tarzia, D.A. An inequality for the coefficient $\sigma$ of the free boundary $s(t)=2\sigma\sqrt{t}$ of the Neumann solution for the two-phase Stefan problem. *Quart. Appl. Math.* (1981) 39: 491-497.


NUMERICAL SOLUTIONS FOR FREE BOUNDARY PROBLEMS WITH EITHER SHRINKAGE OR EXPANSION

MARIELA C. OLGUIN¹, EDUARDO A. SANTILLAN MARCUS² AND MARIA C. SANZIEL³

¹Fac.Cs.Exactas, Ing.y Agrimensura, Universidad Nacional de Rosario
Av.Pellegrini 250-2000-Rosario-Argentina
mcolguin@fceia.unr.edu.ar

²Fac.Cs.Exactas, Ing.y Agrimensura, Universidad Nacional de Rosario
Av.Pellegrini 250-2000-Rosario-Argentina
edus@fceia.unr.edu.ar

³Fac.Cs.Exactas, Ing.y Agrimensura, Universidad Nacional de Rosario
Av.Pellegrini 250-2000-Rosario-Argentina
sanziel@fceia.unr.edu.ar

Key words: Phase change, Two phase Free Boundary Problems, expansion or shrinkage, Finite Difference Method.

Abstract. In this paper, numerical solutions for two-phase Free Boundary problems are obtained. It is supposed that the material in which we attempts to determine the temperature, undergoes processes of expansion or shrinkage. Different boundary conditions are proposed, namely imposed temperature, heat flux or convective condition. For this purpose we develop an algorithm derived from the theoretical analysis of these free boundary problems and a finite difference method with immobilization of the domain. Both procedures are implemented using the Scilab software and the obtained numerical values are compared.

1 INTRODUCTION

Heat transfer problems involving a change of phase due to melting or freezing processes are very important in science and technology [1, 2, 3, 4, 5]. This kind of problems are generally referred as moving-free boundary problems which have been the subject of numerous theoretical, numerical and experimental investigations [6]. Particularly, the density difference between phases has more and more experimental and analytical research interest [7, 8, 9]. In the following we take into consideration a model presented in [10].

We consider a one-dimensional solidification of a pure substance which is initially in liquid state in a bounded interval [0, l]. Initially, the liquid is above the freezing tempera-
ture, and cooling is applied at $x = 0$ while the other end, $x = l$, is kept adiabatic. At the
time $t = 0$, the temperature of the liquid at $x = 0$ comes down to the freezing point and
solidification begins; $x = s(t)$ is the position of the solid-liquid interface. As the liquid
solidifies, it shrinks or expands and appears a region between $x = 0$ and $x = rs(t)$, where
$r = \frac{\rho_1 - \rho_2}{\rho_1} = 1 - \frac{\rho_2}{\rho_1}$ is a parameter, and $\rho_i$ is the density of the region $i$ ($i = 1$: solid;
$i = 2$: liquid). We observe that $r < 1$ and it could be possible for this parameter to take
negative values (if $r \in (0, 1)$ the substance shrinks; if $r < 0$ it expands).

The governing equations are (see [11]):

$$
\alpha_1 \frac{\partial^2 u_1}{\partial x^2}(x, t) = \frac{\partial u_1}{\partial t}(x, t) + r \dot{s}(t) \frac{\partial u_1}{\partial x}(x, t), \quad rs(t) < x < s(t), \quad 0 < t \leq T. \tag{1}
$$

$$
\alpha_2 \frac{\partial^2 u_2}{\partial x^2}(x, t) = \frac{\partial u_2}{\partial t}(x, t), \quad s(t) < x < l, \quad 0 < t \leq T. \tag{2}
$$

where $u_i(x, t)$ is the temperature, $\alpha_i$ is the thermal diffusivity, $k_i$ is the thermal conductivity, $\rho_i$ is the density in region $i$, ($i = 1, 2$), and $r$ is the parameter presented above.

The initial condition is:

$$
u_2(x, 0) = \psi(x), \quad 0 \leq x \leq l, \tag{3}
$$

while the boundary conditions are:

$$
k_1 \frac{\partial u_1}{\partial x}(rs(t), t) = h(u_1(rs(t), t) - u_0), \quad 0 < t \leq T, \tag{4}
$$

$$
\frac{\partial u_2}{\partial x}(l, t) = 0, \quad 0 < t \leq T, \tag{5}
$$

the conditions at the free unknown boundary $x = s(t)$ are:

$$
u_1(s(t), t) = u_2(s(t), t) = u^*, \quad 0 < t \leq T, \tag{6}
$$

$$
k_1 u_{1x}(s(t), t) - k_2 u_{2x}(s(t), t) = \rho_1 L \dot{s}(t), \quad 0 < t \leq T, \tag{7}
$$

where $h$ is the heat transfer coefficient, $u_0$ is the external temperature, $u^*$ is the freezing
temperature ($u_0 < u^*$) and $L$ is the latent heat of fusion by unit of mass.

The goal of this paper is to obtain numerical solutions for free boundary problems in
which shrinkage or expansion occurs.


2 ANALYTICAL RESULTS

In [10] it was proved that under appropriate assumptions there exists a unique solution \( \{u_1(x, t), u_2(x, t), s(t)\} \) to the problem (1)-(7).

In [12] the explicit solutions for three free boundary problems were given through a representation of the similarity type. The problems considered in this paper are alike problem (1)-(7), but they considered a semi-infinite domain with three different boundary conditions on the left boundary, instead of (4), namely:

- a temperature condition:
  \[ u_1(rs(t), t) = A; \]  

- a heat-flux condition:
  \[ k_1 \frac{\partial u_1}{\partial x}(rs(t), t) = \frac{q_0}{\sqrt{t}}; \]  

- a convective cooling condition:
  \[ k_1 \frac{\partial u_1}{\partial x}(rs(t), t) = \frac{h_0}{\sqrt{t}} (u_1(rs(t), t) - u_0), \]

The explicit solutions are [12]:

- in the case of a temperature condition:
  \[ u_1(x, t) = A + (u^* - A) \frac{erf \left( \sqrt{\frac{\alpha_2}{\alpha_1}} \lambda_1 \left( \frac{x}{2\lambda_1 \sqrt{\alpha_2 t}} - r \right) \right)}{erf \left( \sqrt{\frac{\alpha_2}{\alpha_1}} \lambda_1 (1 - r) \right)}, \quad rs(t) < x < s(t), \quad t > 0 \]  

  \[ u_2(x, t) = B - (B - u^*) \frac{erfc \left( \frac{x}{2\sqrt{\alpha_2 t}} \right)}{erfc(\lambda)} \]  

- in the case of a heat-flux condition:
  \[ u_1(x, t) = u^* - \frac{q_0 \sqrt{\alpha_1}}{k_1} \left( erf \left( \sqrt{\frac{\alpha_2}{\alpha_1}} \lambda_2 (1 - r) \right) - erf \left( \sqrt{\frac{\alpha_2}{\alpha_1}} \lambda_2 \left( \frac{x}{2\lambda_2 \sqrt{\alpha_2 t}} - r \right) \right) \right), \quad rs(t) < x < s(t), \quad t > 0 \]  

  \[ u_2(x, t) = B - (B - u^*) \frac{erfc \left( \frac{x}{2\sqrt{\alpha_2 t}} \right)}{erfc(\lambda_2)} \]  

\[ x > s(t), \quad t > 0 \]
• in the case of a convective cooling condition:

\[ u_1(x, t) = u_0 + (u^* - u_0) \frac{k_1 + \sqrt{\pi \alpha_1} h_0 \text{erf} \left( \frac{x}{2\sqrt{\alpha_1}} - r\lambda_3 \sqrt{\frac{\alpha_2}{\alpha_1}} \right)}{k_1 + \sqrt{\pi \alpha_1} h_0 \text{erf} \left( \frac{\sqrt{\alpha_2}}{\alpha_1} \lambda_3 (1 - r) \right)}, \quad rs(t) < x < s(t), \quad t > 0 \]

(15)

\[ u_2(x, t) = B - (B - u^*) \frac{\text{erfc} \left( \frac{x}{2\sqrt{\alpha_2}} \right)}{\text{erfc} (\lambda_3)}, \quad x > s(t), \quad t > 0 \]

(16)

and the free boundaries \( s(t) \) in the three cases are of the type:

\[ s(t) = 2\lambda_i \sqrt{\alpha_2 t}; \quad t > 0 \]

(17)

where the coefficients \( \lambda_i \) (\( i = 1, 2, 3 \)), are the unique solutions of three different equations.

3 NUMERICAL APPROACH

3.1 Implicit finite difference method with immobilization of the domain

In order to solve numerically the free boundary problem (1)-(7), first we fix the moving front by a suitable choice of new space coordinates. Following Landau’s coordinate transformation [13, 14], we define:

\[ \xi_1 = \frac{x}{s(t)}, \quad \text{for} \quad rs(t) < x < s(t) \quad \text{and} \quad \xi_2 = \frac{x - s(t)}{l - s(t)} \quad \text{for} \quad s(t) < x < l. \]

The new unknown functions \( U_1(\xi_1, t) \) and \( U_2(\xi_2, t) \) must satisfy:

\[ \frac{\partial U_1}{\partial t}(\xi_1, t) = \frac{\alpha_1}{s^2(t)} \frac{\partial^2 U_1}{\partial \xi_1^2}(\xi_1, t) + (\xi_1 - r) \frac{\dot{s}(t)}{s(t)} \frac{\partial U_1}{\partial \xi_1}(\xi_1, t), \quad r < \xi_1 < 1, \quad 0 < t \leq T \]

(18)

\[ \frac{\partial U_2}{\partial t}(\xi_2, t) = \frac{\alpha_2}{(l - s(t))^2} \frac{\partial^2 U_2}{\partial \xi_2^2}(\xi_2, t) + (1 - \xi_2 - \frac{1}{l - s(t)} - 1) \frac{\dot{s}(t)}{s(t)} \frac{\partial U_2}{\partial \xi_2}(\xi_2, t), \quad 0 < \xi_2 < 1, \quad 0 < t \leq T \]

(19)

\[ U_2(\xi_2, 0) = \psi(\xi_2), \quad 0 < \xi_2 < 1 \]

(20)

\[ k_1 \frac{\partial U_1}{\partial \xi_1}(r, t) = h s(t) (U_1(r, t) - u_0), \quad 0 < t \leq T \]

(21)

\[ U_1(1, t) = U_2(0, t) = u^*, \quad 0 < t \leq T \]

(22)

\[ \frac{\partial U_2}{\partial \xi_2}(1, t) = 0, \quad 0 < t \leq T \]

(23)
\[
\dot{s}(t) = \frac{1}{\mu_1 L} \left( \frac{k_1 \partial U_1}{s(t) \partial \xi_1}(1, t) - \frac{k_2}{l - s(t)} \frac{\partial U_2}{\partial \xi_2}(0, t) \right), \quad 0 < t \leq T. \tag{24}
\]

Then we use an Implicit Finite Difference Method (IFDM). We consider an uniform partition of the domains \([r, 1]\) and \([0, 1]\) in \(N_i\) subintervals \((i = 1, 2)\). We call \(U^i_{k, j}\) the approximate values of the unknown functions \(U_i(x, t_k)\) and \(s_k\) the approximate value of \(s(t_k)\), \((i = 1, 2; j = 1, \ldots, N_i + 1; k = 1, 2, \ldots, k_{\text{max}})\).

Using the usual implicit finite-difference discretization in time, combined with central differences for space derivatives, we get a tridiagonal system for each phase and for each time \(t_k = k \Delta t\) (a positive fixed time step), from equations (18)-(23):

1. In the solid phase:
   \[
   A^k U^1_k = T^k
   \]
   where:
   \[
   U^1_k = (U_1^{1, k}, \ldots, U_1^{N_1, k})^T
   \]
   \[
   A^k_{1,1} = 1 + h s_k \frac{\Delta \xi_1}{k_1}, \quad A^k_{1,2} = -1,
   \]
   for \(j = 2, \ldots, N_1 - 1\):
   \[
   A^k_{j,j-1} = -a^k_1, \quad A^k_{j,j} = 1 + 2a^k_1 + a^k_2, \quad A^k_{j,j+1} = -(a^k_1 + a^k_2),
   \]
   and \(A^k_{N_1, N_1 - 1} = -a^k_1, \quad A^k_{N_1, N_1} = 1 + 2a^k_1 + a^k_2, \)
   with:
   \[
   a^k_1 = \frac{\alpha_1 \Delta t}{s_k \Delta \xi_1}, \quad a^k_2 = \left(1 - \frac{s_k-1}{s_k}\right) \left(\frac{\xi_1(j)-r}{\Delta \xi_1}\right)
   \]
   for \(j = 1, \ldots, N_1\);
   \[
   T^k_1 = h s_k \frac{\Delta \xi_1}{k_1}, \quad T^k_j = U^1_{k,j-1}, \quad j = 2, \ldots, N_1 - 1, \quad T^k_{N_1} = U^1_{k,N_1-1} + (a^k_1 + a^k_2) u^*;
   \]
   \[
   U^1_{k,N_1+1} = u^*.
   \]

2. In the liquid phase:
   \[
   B^k U^2_k = S^k
   \]
   where:
   \[
   U^2_k = (U_2^{2, k}, \ldots, U_2^{N_2+1, k})^T
   \]
   \[
   B^k_{1,1} = 1 + 2a^k_3 + a^k_4, \quad B^k_{1,2} = -a^k_3,
   \]
   for \(j = 2, \ldots, N_2 - 1\):
   \[
   B^k_{j,j-1} = -(a^k_3 + a^k_4), \quad B^k_{j,j} = 1 + 2a^k_3 + a^k_4, \quad B^k_{j,j+1} = -a^k_3,
   \]
   \[
   B^k_{N_2,N_2-1} = 1, \quad B^k_{N_2,N_2} = -1
   \]
   with:
   \[
   a^k_3 = \frac{\alpha_2 \Delta t}{(l-s_k) \Delta \xi_2}, \quad a^k_4 = \frac{(s_k-s_k-1)(1-\xi_2(j))}{(l-s_k) \Delta \xi_2}
   \]
   \[
   S^k_1 = U^2_{k,1} + \frac{\Delta t a^k_3 u^*}{(l-s_k) \Delta \xi_2}, \quad S^k_j = U^2_{k,j+1,k-1}, \quad j = 2, \ldots, N_2 - 1, \quad S^k_{N_2} = 0;
   \]
   and \(U^1_{k,1} = u^*\).
From the equation (24), through an explicit discretization in time, we can obtain $s_k$ as:

$$s_k = s_{k-1} + \frac{\Delta t}{\rho_1 L} \left( \frac{k_1 U_1^{N_1+1,k-1} - U_1^{N_1,k-1}}{s_{k-1} - 1} - \frac{k_2 U_2^{2,k-1} - U_2^{1,k-1}}{l - s_{k-1}} \right).$$

We also solve numerically the cases with temperature condition and with flux condition on the left side, i.e.:

$$U_1(r, t) = \phi(t), \quad \text{or} \quad k_1 \frac{\partial U_1(r, t)}{\partial \xi_1} = \phi(t), \quad 0 < t \leq T,$$

instead of (21). In these cases we obtain similar tridiagonal systems.

### 3.2 Numerical Experiments

For the numerical experiments we consider two materials, water and aluminum. So we have an expansion model and a shrinkage model respectively. In Table 1 we show the values of the thermal coefficients for the two materials [15, 16].

<table>
<thead>
<tr>
<th></th>
<th>Water</th>
<th>Aluminum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho_1 \left[ \frac{kg}{m^3} \right]$</td>
<td>917</td>
<td>2707</td>
</tr>
<tr>
<td>$\rho_2 \left[ \frac{kg}{m^3} \right]$</td>
<td>1000</td>
<td>2528</td>
</tr>
<tr>
<td>$c_1 \left[ \frac{J}{kg \cdot ^\circ C} \right]$</td>
<td>2114</td>
<td>905</td>
</tr>
<tr>
<td>$c_2 \left[ \frac{J}{kg \cdot ^\circ C} \right]$</td>
<td>4181</td>
<td>1200</td>
</tr>
<tr>
<td>$k_1 \left[ \frac{W}{m \cdot ^\circ C} \right]$</td>
<td>2</td>
<td>237</td>
</tr>
<tr>
<td>$k_2 \left[ \frac{W}{m \cdot ^\circ C} \right]$</td>
<td>0.613</td>
<td>234</td>
</tr>
<tr>
<td>$L \left[ \frac{J}{kg} \right]$</td>
<td>$3 \times 10^5$</td>
<td>$394 \times 10^4$</td>
</tr>
<tr>
<td>$u^* \left[ ^\circ C \right]$</td>
<td>0</td>
<td>658</td>
</tr>
</tbody>
</table>

The explicit solutions mentioned above (15),(16),(17) were used to compare the numerical solutions and also to initialize the numerical schemes.

In Table 2 we show the relative errors when we consider increasing values of $N_1$, for the case of water with a convective cooling condition.

In figures 1 and 2 we can see respectively, the free boundaries positions and the temperature distribution at final time $T = 5h 30'$ for the case of water with convective left boundary condition. We have considered $u_0 = -5^\circ C$, $\psi(x) = 20^\circ C$ and $h = 2.53 \times 10^6 \frac{W}{m^2 \cdot ^\circ C}$.
Table 2: Relative Errors

<table>
<thead>
<tr>
<th>$N_1$</th>
<th>for the free boundary</th>
<th>for the temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.929634 $10^{-3}$</td>
<td>0.36642064 $10^{-8}$</td>
</tr>
<tr>
<td>200</td>
<td>0.929247 $10^{-3}$</td>
<td>0.36642058 $10^{-8}$</td>
</tr>
<tr>
<td>400</td>
<td>0.929056 $10^{-3}$</td>
<td>0.36642055 $10^{-8}$</td>
</tr>
<tr>
<td>500</td>
<td>0.929018 $10^{-3}$</td>
<td>0.36642055 $10^{-8}$</td>
</tr>
</tbody>
</table>

In figures 3 and 4 we can see the free boundaries positions and the temperature distribution at final time $T = 5h 30'$ for the case of aluminum with convective left boundary condition. We have considered $u_0 = -5 \, ^\circ C$, $\psi(x) = 658 \, ^\circ C$ and $h = 6.013 \times 10^9 \frac{W}{m^2 \, ^\circ C}$.

4 CONCLUSIONS

Heat transfer problems involving a change of phase are very important in technology. The difference in densities between phases is of particular interest in a great number of processes. We analyze two-phase free boundary problems with shrinkage or expansion of the material because of the difference in densities.

In [10] it was considered the case with convective cooling condition on the left side and it was proved the existence of unique solution for the free-boundary problem. In [12] they found explicit solutions but only for the cases of constant temperature condition or a special type of heat-flux condition or a special type of convective condition on the left side.

In this paper we have developed and we have implemented, using the free software Scilab, a numerical method which allows to obtain the numerical values of the temperature distribution and the position of the free boundary for this type of problems, with different boundary conditions at the left boundary: temperature (not necessarily constant), heat-flux or convective condition.

The results obtained from the IFDM are in good agreement with those of the explicit solutions [12], and exhibit the expected convergence as the mesh size is refined.

5 ACKNOWLEDGEMENTS

This paper has been partially supported by Project PID ING403, Tratamiento Numérico y Problemas de Control Óptimo vinculados a Problemas de Frontera Libre (Universidad Nacional de Rosario).

REFERENCES


Figure 1: Free boundaries for water

Figure 2: Temperature distributions for water at $T = 5h 30'$
**Figure 3:** Free boundaries for aluminum

**Figure 4:** Temperature distributions for aluminum at $T = 5h30'$
DETERMINATION OF TWO UNKNOWN THERMAL COEFFICIENTS THROUGH A MUSHY ZONE WITH A CONVECTIVE OVERSPECIFIED BOUNDARY CONDITION

ANDREA N. CERETANI AND DOMINGO A. TARZIA

CONICET - Departamento de Matemática, Fac. de Ciencias Empresariales, Univ. Austral Paraguay 1950, S2000FZF Rosario, Argentina
aceretani@austral.edu.ar, dtarzia@austral.edu.ar

Key words: Phase Change, Convective Condition, Lamé-Clapeyron-Stefan Problem, Mushy Zone, Solomon-Wilson-Alexiades Model, Unknown Thermal Coefficients

Abstract. We consider a semi-infinite material characterized by $x > 0$ which is initially assumed to be liquid at its melting temperature. At time $t = 0$ a heat flux condition is imposed at the fixed face $x = 0$, and a solidification process with a mushy zone begins. We impose an overspecified convective condition at $x = 0$ with the aim of the simultaneous determination of two thermal coefficients among $l$ (latent heat by unit mass), $k$ (thermal conductivity), $\rho$ (mass density), $c$ (specific heat), $\epsilon$ and $\gamma$ (coefficients that characterize the mushy zone), when $q_0 > 0$, $h_0 > 0$ (coefficients that characterize the heat flux and heat transfer at $x = 0$, respectively), $D_{\infty}$ (external temperature at $x = 0$) and one of the two boundaries of the mushy zone are determined experimentally. This lead us to the study of 15 different cases. We present the study of some of them, besides explicit formulae for the unknown thermal coefficients.

1 Introduction

Heat transfer problems with a phase-change such as melting and freezing have been studied in the last century due to their wide scientific and technological applications. Some previous books in the subject are [1–8, 14].

In this paper we consider a semi-infinite material characterized by $x > 0$ that is initially assumed to be liquid at its melting temperature, which without loss of generality we assume equal to $0^\circ$C. At time $t = 0$ a flux condition is imposed at the fixed face $x = 0$, and a solidification process begins where the following three regions can be distinguished [9,11]:

1. liquid region at temperature $T(x, t) = 0$: $D_l = \{(x, t) \in \mathbb{R}^2 / x > r(t), \ t > 0\}$,
2. solid region at temperature $T(x, t) < 0$: $D_s = \{(x, t) \in \mathbb{R}^2 / 0 < x < s(t), \ t > 0\}$,
3. mushy region at temperature $T(x, t) = 0$: $D_p = \{(x, t) \in \mathbb{R}^2/ s(t) < x < r(t), \ t > 0\}$; being $s = s(t)$ and $r = r(t)$ the functions that characterize the free boundaries of the mushy zone. We make the following assumptions on the structure of the mushy zone, which we consider isothermal:

1. the material contains a fixed portion $\epsilon l$ of the total latent heat per unit mass $l > 0$, with $0 < \epsilon < 1$, that is:

$$kT_x(s(t), t) = \rho l [\epsilon \dot{s}(t) + (1 - \epsilon) \dot{r}(t)], \ t > 0,$$

where $k > 0$ is the thermal conductivity and $\rho > 0$ is the density mass of the material,

2. its width is inversely proportional to the gradient of temperature, that is:

$$T_x(s(t), t)(r(t) - s(t)) = \gamma, \ t > 0$$

where $\gamma > 0$.

Encouraged by the recent works [12, 13] and with the aim of the simultaneous determination of the temperature $T = T(x, t)$, the free boundary $x = r(t)$ and two of the thermal coefficients among $l$ (latent heat by unit mass), $k$ (thermal conductivity), $\rho$ (mass density), $c$ (specific heat), $\epsilon$ and $\gamma$ (coefficients which characterize the mushy zone), we impose an overspecified convective condition at $x = 0$ (see condition (7) below), which leads us to the following free boundary problem:

$$\rho c T_t(x, t) - kT_{xx}(x, t) = 0 \quad 0 < x < s(t), t > 0 \quad (1)$$

$$T(s(t), t) = 0 \quad t > 0 \quad (2)$$

$$kT_x(s(t), t) = \rho l [\epsilon \dot{s}(t) + (1 - \epsilon) \dot{r}(t)] \quad t > 0 \quad (3)$$

$$T_x(s(t), t)(r(t) - s(t)) = \gamma \quad t > 0 \quad (4)$$

$$r(0) = s(0) = 0 \quad (5)$$

$$kT_x(0, t) = \frac{q_0}{\sqrt{t}} \quad t > 0 \quad (6)$$

$$kT_x(0, t) = \frac{h_0}{\sqrt{t}} (T(0, t) + D_\infty) \quad t > 0 \quad (7)$$

where $q_0 > 0$ and $h_0 > 0$ are the coefficients that characterize the heat flux and the heat transfer at $x = 0$, respectively, and $-D_\infty < 0$ is the external temperature at $x = 0$.

We assume the free boundary $s(t)$ is given by:

$$s(t) = 2\sigma \sqrt{t}, \quad t > 0 \quad (\sigma > 0) \quad (8)$$

and the coefficients $q_0$, $h_0$, $D_\infty$ and $\sigma$ are determined experimentally.

The determination of the two unknown thermal coefficients for the one-phase Lamé-Clapeyron-Stefan problem without a mushy zone was done in [10]. The goal of this paper is to obtain the explicit solution to problem (1)-(7) with two unknown thermal coefficients and the necessary and sufficient conditions on data in order to obtain the explicit formulae for the two unknown thermal coefficients.
2 Solution of the Problem

We have:

**Theorem 1.** The solution to problem (1)-(7) with \( x = s(t) \) as in (8) is given by:

\[
T(x, t) = -\frac{q_0 \sqrt{\alpha t}}{k} \operatorname{erf} \left( \frac{\sigma}{\sqrt{\alpha t}} \right) \left[ 1 - \frac{\operatorname{erf} \left( \frac{x}{2\sqrt{\alpha t}} \right)}{\operatorname{erf} \left( \frac{\sigma}{\sqrt{\alpha t}} \right)} \right] \quad 0 < x < s(t), t > 0
\]

\[
r(t) = \left( \frac{\gamma k \exp(\sigma^2/\alpha)}{q_0} + 2\sigma \right) \sqrt{t} \quad t > 0
\]

if and only if the parameters involved in the problem (1)-(7) satisfy the following two equations:

\[
\frac{q_0}{\rho l} = \left[ \sigma + \gamma k (1 - \epsilon) \exp(\sigma^2/\alpha) \right] \exp(\sigma^2/\alpha)
\]

\[
\operatorname{erf} \left( \frac{\sigma}{\sqrt{\alpha}} \right) = \frac{k D_\infty}{q_0 \sqrt{\alpha \pi}} \left( 1 - \frac{q_0}{h_0 D_\infty} \right)
\]

**Proof.** Since the solution to problem (1)-(7) with \( x = s(t) \) as in (8) has the form [9,11,13]:

\[
T(x, t) = A + B \operatorname{erf} \left( \frac{x}{2\sqrt{\alpha t}} \right) \quad 0 < x < s(t), t > 0
\]

\[
r(t) = 2\mu \sqrt{\alpha l} \quad t > 0
\]

where \( \alpha = \frac{k}{\rho c} \) (thermal diffusivity), by imposing conditions (2)-(7) we obtain that the coefficients \( A, B \) and \( \mu \) must be given by:

\[
A = -\frac{q_0 \sqrt{\alpha \pi}}{k} \operatorname{erf} \left( \frac{\sigma}{\sqrt{\alpha}} \right), \quad B = \frac{q_0 \sqrt{\alpha \pi}}{k} \quad \text{and} \quad \mu = \frac{\gamma k \exp(\sigma^2/\alpha)}{2q_0 \sqrt{\alpha}} + \frac{\sigma}{\sqrt{\alpha}}
\]

that is, the solution to problem (1)-(7) is given by (13)-(14), and that the parameters involved in the problem must satisfies equations (11) and (12).

The problem of the determination of the temperature \( T = T(x, t) \), the free boundary \( x = r(t) \) and two coefficients among \( l, \gamma, \epsilon, k, \rho \) and \( c \), leads us to the study of 15 different cases, which we classify as:

- Case 1: Determination of \( \epsilon \) and \( \gamma \)
- Caso 2: Determination of \( \epsilon \) and \( l \)
- Caso 3: Determination of \( \gamma \) and \( l \)
- Case 4: Determination of \( \epsilon \) and \( k \)
- Caso 5: Determination of \( \epsilon \) and \( \rho \)
- Caso 6: Determination of \( \epsilon \) and \( c \)
Case 7: Determination of $\gamma$ and $k$, Caso 8: Determination $\gamma$ and $\rho$,

Case 9: Determination of $\gamma$ and $c$, Case 10: Determination of $l$ and $k$,

Case 11: Determination of $l$ and $\rho$, Caso 12: Determination of $l$ and $c$,

Case 13: Determination of $k$ and $\rho$, Caso 14: Determination of $k$ and $c$,

and

Caso 15: Determination of $\rho$ and $c$.

We will present the study of some of them in the following Section.

3 Results

We consider the cases 7 and 13.

Theorem 2 (Case 7: determination of $\gamma$ and $k$). If in problem (1)-(7) we consider $x = s(t)$ as in (8) and the thermal parameters $\gamma$ and $k$ as unknowns, then its solution is given by (13)-(14) with $\gamma$ and $k$ given by:

$$\gamma = \frac{2q_0\xi^2}{\sigma \rho c} \left( \frac{q_0}{\sigma \rho l} \exp(-\xi^2) - 1 \right) \exp(-\xi^2) > 0$$ (16)

$$k = \rho c \left( \frac{\sigma}{\xi} \right)^2 > 0$$ (17)

being $\xi$ the unique positive solution of the equation:

$$f(x) = \frac{\sigma \rho c D_{\infty}}{q_0 \sqrt{\pi}} \left( 1 - \frac{q_0}{h_0 D_{\infty}} \right), \quad x > 0,$$ (18)

where $f$ is the function defined by:

$$f(x) = x \ \text{erf}(x), \quad x > 0,$$ (19)

if and only if the parameters $q_0$, $h_0$, $D_{\infty}$, $\sigma$, $l$, $\rho$ and $c$ satisfy the following three inequalities:

$$\frac{q_0}{\sigma \rho l} - 1 > 0$$ (20)

$$0 < 1 - \frac{q_0}{h_0 D_{\infty}} < \frac{q_0 \sqrt{\pi}}{\sigma \rho c D_{\infty}} \sqrt{\ln \left( \frac{q_0}{\sigma \rho l} \right)} \ \text{erf} \left( \sqrt{\ln \left( \frac{q_0}{\sigma \rho l} \right)} \right)$$ (21)
Proof. We know from Theorem 1 that (13)-(14) is the solution to problem (1)-(7) with \( x = s(t) \) as in (8) if and only if \( \gamma \) and \( k \) satisfy the system of equations (11)-(12). This system can be written as:

\[
\gamma = \frac{2q_0\xi^2}{\sigma pc} \left( \frac{q_0}{\sigma \rho l} \exp (-\xi^2) - 1 \right) \exp (-\xi^2) \tag{22}
\]

\[
f(\xi) = \frac{\sigma pc D_\infty}{q_0\sqrt{\pi}} \left( 1 - \frac{q_0}{h_0 D_\infty} \right) \tag{23}
\]

where

\[
\xi = \frac{\sigma}{\sqrt{\alpha}} = \sigma \sqrt{\frac{pc}{k}}. \tag{24}
\]

Since \( f \) is a strictly increasing function in \( \mathbb{R}^+ \) such that \( f(0^+) = 0 \) and \( f(+\infty) = +\infty \), we have that equation (23) admits a unique positive solution if and only if the first inequality in (21) holds. In other words, that equation (12) admits a unique solution \( k \), which is given by (17) (see (24)) where \( \xi \) is the unique positive solution of the equation (18), if and only if the first inequality in (21) holds. Finally, let us observe that the coefficient \( \gamma \) given in (22) is positive if and only if \( \frac{q_0}{\sigma \rho l} \exp (-\xi^2) > 1 \), that is, if and only if inequality (20) holds and

\[
\xi < \sqrt{\ln \left( \frac{q_0}{\sigma \rho l} \right)}. \tag{25}
\]

By applying the function \( f \) side by side of this last inequality and taking into account that \( f \) is strictly increasing in \( \mathbb{R}^+ \) and \( \xi \) is the unique solution of the equation (18), it follows that this last inequality is equivalent to the second inequality in (21). \( \square \)

**Theorem 3** (Case 13: determination of \( k \) and \( \rho \)). If in problem (1)-(7) we consider \( x = s(t) \) as in (8) and the thermal parameters \( k \) and \( \rho \) as unknowns, then its solution is given by (13)-(14) with \( k \) and \( \rho \) given by:

\[
k = \frac{q_0\sigma \sqrt{\pi}}{D_\infty \left( 1 - \frac{q_0}{h_0 D_\infty} \right)} \frac{\text{erf}(\xi)}{\xi} > 0 \tag{26}
\]

\[
\rho = \frac{q_0\sqrt{\pi}}{c\sigma D_\infty} \xi \text{ erf} (\xi) \tag{27}
\]

being \( \xi \) the unique positive solution of the equation:

\[
g(x) = h(x), \quad x > 0, \tag{28}
\]

where functions \( g \) and \( h \) are defined by:

\[
g(x) = \frac{a}{\exp (x^2) \text{ erf}(x)} \quad \text{and} \quad h(x) = bx + c \exp (x^2) \text{ erf}(x), \quad x > 0 \tag{29}
\]
with
\[ a = \frac{2cD_\infty}{l\sqrt{\pi}} \left( 1 - \frac{q_0}{h_0D_\infty} \right)^2 > 0, \quad b = 2 \left( 1 - \frac{q_0}{h_0D_\infty} \right) > 0, \quad c = \frac{\gamma \sqrt{\pi}(1 - \epsilon)}{D_\infty} > 0, \]

if and only if the parameters \( q_0, h_0 \) and \( D_\infty \) satisfy the following inequality:
\[ 1 - \frac{q_0}{h_0D_\infty} > 0 \] \hspace{1cm} (30)

**Proof.** We know from Theorem 1 that (13)-(14) is the solution to problem (1)-(7) with \( x = s(t) \) as in (8) if and only if \( k \) and \( \rho \) satisfy the system of equations given by (11)-(12). This system can be written as:
\[ \frac{q_0c\sigma^2}{lk^2} = \left[ \sigma + \frac{\gamma k(1 - \epsilon)}{2q_0} \exp (\xi^2) \right] \exp (\xi^2) \] \hspace{1cm} (32)
\[ \text{erf}(\xi) = \frac{kD_\infty}{q_0\sigma\sqrt{\pi}} \left( 1 - \frac{q_0}{h_0D_\infty} \right) \] \hspace{1cm} (33)

where \( \xi \) is as in (24). Let us observe that inequality (31) is a necessary condition for the existence of a solution \( k > 0 \) and \( \rho > 0 \) to system (32)-(33). Then, henceforth, we assume that inequality (31) holds.

From equation (33), we have that \( k \) is given by (26). By replacing this expression for \( k \) in (32), we obtain that equation (32) is equivalent to \( g(\xi) = h(\xi) \), that is \( \xi \) must satisfy equation (28).

Since \( g \) is a strictly decreasing function in \( \mathbb{R}^+ \) such that \( g(0^+) = +\infty \) and \( g(+\infty) = 0 \), and \( h \) is a strictly increasing function such that \( h(0^+) = 0 \) and \( h(+\infty) = +\infty \), it follows that there exists a unique positive solution to equation (28). Therefore, condition (31) is a necessary and sufficient condition for the existence and uniqueness of the solution to system (11)-(12), which is given by (26) and (27) (see (24)), where \( \xi \) is the unique positive solution to equation (28).

**Acknowledgements**

This paper was partially supported by the Projects PIP N° 0534 of CONICET - Universidad Austral, Rosario, Argentina and AFOSR-SOARD Grant FA 9550-14-1-0122.

**REFERENCES**


CUMULATIVE NUTRIENT UPTAKE BY ROOTS OF CROPS AS SIMULATED BY FIXED AND MOVING BOUNDARY MODELS. CORRECTIONS AND IMPROVEMENTS

JUAN C. REGINATO¹, JORGE L. BLENGINO¹ AND DOMINGO A. TARZIA²

¹ Departamento de Física, Facultad de Ciencias Exactas, Universidad Nacional de Río Cuarto. Ruta Nacional 36, Km 601, X5804BYA Río Cuarto, Córdoba, Argentina, jreginato@exa.unrc.edu.ar

² Departamento de Matemática - CONICET, Facultad de Ciencias Empresariales, Universidad Austral Paraguay 1950, S2000FZF Rosario, ARGENTINA, DTarzia@austral.edu.ar

Key words: Nutrient Uptake by Roots, Moving Boundary Problems, FEM.

Summary This work examines the relevance of physical models used to study the flux and nutrient uptake by roots of crops. The physical models studied are the one-dimensional fixed boundary model of Cushman-Barber [1] and an improved version of our one-dimensional moving boundary model [2]. The moving boundary model is solved by immobilizing the domain in dimensionless variables and computed by the finite elements method. To estimate the cumulative nutrient uptake a generalized and verified formula is used for both models. For simulations of nutrient uptake three datum sets extracted from literature were used. First, we compute the cumulative uptake of nutrient of low mobility as K and P by pine seedling in high soil concentrations. Second, we compute the K uptake by maize, wheat and sugar beet for low K soil and soil with K addition and finally we compute the P uptake by wheat to low concentrations. Third, we compute the P uptake for peanut for low, intermediate and high soil concentrations. For low concentrations, large variations of root density and low numbers of Peclet the moving boundary produces better predictions particularly for K. For P the moving boundary produces better predictions only at low concentrations being these predictions comparable to the obtained by 3D-dimensional architectural models [3]. Finally, in the light of these findings, conclusions drawn by previous papers [4] could be reinterpreted.

1 INTRODUCTION

Over the past four decades different mechanistic nutrient uptake models have been developed to simulate nutrient uptake. Two categories of models have evolved: steady state and transient models. NUTRIENT UPTAKE [6] and NST 3.0 [7] are examples of a transient model with a numerical solution, while SSAND [8] and PCATS [9] are steady state models. The Barber-Cushman model [1] is a well-known and widely-used model in this category. The model treats the system as two concentric cylinders, where the inner is the root (with constant radius, and no extensions like branching, lateral roots, root hairs or mycorrhizal hypha), whose center is the spatial reference to the soil-root system, with radial orientation. The soil, (assumed homogeneous and isomorphic, with constant moisture content) forms an external
cylinder around the root, also with a constant radius. Movement of water and solutes in the soil system is radial to the root only, by mass-flow and diffusion, following Nye and Marriott 1969 [5]. Water flow, controlled by the transpiration demand (assumed constant with time), obeys the radial geometry of the system and mass conservation. Nutrient uptake rate is a function of concentration of the ion in question in the soil solution at the root surface, assuming that uptake occurs from a solution only, without interaction with other solutes. Updates to this basic feature include moving boundaries, the external radius (the available soil extent to each root) to account for root growth with time and consequent increase in root density [2]. NUTRIENT UPTAKE model and NST 3.0 are the personal computer version of the Barber-Cushman model. Further refinements of the Nye and Marriott derived models consisted to upscale from the root segment to the whole root system, and accounted for root growth. By using up scaling Roose et al. [10,11] provide a fully explicit ‘approximate’ analytical solution to the Nye–Tinker–Barber model and applied this solution to more complex root branching structure. Roose et al. showed that the method used to upscale may lead to substantial differences in the predicted uptake of nutrients between their models and NST 3.0 model. When applied to nutrients such as K and P, such models have generally proved quite efficient at predicting the acquisition over time scales of days or weeks for soils receiving high K or P inputs, but almost systematically failed in low input conditions [12,13]. Under such conditions, those models actually underestimate the observed uptake flux, which suggests that other processes than those accounted for by the models could be operating, and ultimately driving nutrient acquisition. However, a comparison of nutrient uptake predictions against experimentally measured values showed that the last version of three process-based models (NST 3.0, SSAND, and PCATS) largely underestimated P uptake for three woody plant species. This pattern showed that including mycorrhizal uptake in the simulations was not sufficient to predict accurately nutrient uptake under the low nutrient concentrations. These results suggested that rhizospheric effects, not yet taken into account in these models, could be carried out to improve their predictive ability. Further 3D root system architecture models were RootTyp, SimRoot, ROOTMAP, SPACSYS, R-SWMS, and RootBox [14] and they are being used to study how specific root traits affect the uptake of a variety of soil resources such as nitrogen, phosphorus, and water. From other point of view, heat and mass transfer with phase change problems such as evaporation, condensation, freezing, melting, sublimation, have wide application in separation processes, food technology, heat and mixture migration in soils and grounds, etc. This kind of problem are known in the literature as free or moving boundary problems depending if the interface is unknown or known a priori. A large bibliography and a review of explicit solutions on free and moving boundary problems for the heat-diffusion equation were given in [15].The methodology of free and moving boundary problems have been also applied to agronomic problems. Thus, there are previous papers in this area for which the nutrient uptake has been implicitly modeled by moving boundary problems, for example, Abbes et al. [16], Huguenin and Kirk, [17]. Explicit one-dimensional moving boundary model applied to root growth and nutrient uptake was presented in Reginato et al.[2], Jonard et al.[18]. The goal of this paper is to consider an improved version of the moving boundary model applied to uptake of ions of low, medium and high availability by roots. In particular, we will revise and compare the uptake of ions through model NST 3.0, and the moving boundary model. For both models, we use a new generalized cumulative uptake formula and the moving boundary model is solved by the adaptive finite elements
method. Moreover, we also compare the cumulative uptake predicted by a more complex root branching system model with ones obtained by our moving boundary model.

2 THE MOVING BOUNDARY MODEL

This model is based on the same assumptions formulated by the Barber-Cushman model but, now, the model incorporates a new boundary condition for root competition (among roots of root system) which represents the net flux on the moving boundary R(t). This moving boundary is given by the instantaneous half distance between roots axis which is the result of the root length variation. Thus, moving boundary R(t) is a function of the instantaneous root length \( \ell (t) \) which is a known function of time. A representation of the new condition can be visualised assuming a fixed volume of soil in which the root system is distributed like a homogenous piling up by roots, i.e., we propose an idealized total root system submerged in a fixed volume of soil (pots) instead of a single root in an infinite volume of soil. The conditions of humidity, light and temperature are assumed to be controlled (as in a growth chamber). Based on these assumptions and using root length density as a function of \( t \), \( R(t) \) (the moving boundary), the following set of equations and boundary conditions in cylindrical coordinates are used:

\[
(\phi + b) \frac{\partial C}{\partial t} = \frac{\phi D}{r} \left( \frac{\partial}{\partial r} \left( r \frac{\partial C}{\partial r} \right) \right) + \frac{v_s}{r} \frac{\partial C}{\partial r}, \quad s_o < r < R(t), \quad t > 0
\]

(1)

\[
C(r,0) = C_o(r), \quad s_o < r < R_o
\]

(2)

\[
D \frac{\partial C(s_o,t)}{\partial r} + vC(s_o,t) = \frac{J_m}{K_m} \left[ C(s_o,t) - C_u \right], \quad t > 0
\]

(3)

\[
D \frac{\partial C(R(t),t)}{\partial r} + \frac{v_s}{R(t)} C(R(t),t) = 0, \quad t > 0
\]

(4)

where the moving boundary is given by:

\[
R(t) = \sqrt{\frac{\ell_o}{\ell(t)}} \left( R_o^2 - s_o^2 \right) + s_o, \quad t > 0
\]

(5)

where \( r \) is the radial distance from the axis of the root [cm]; \( t \) is the time [s]; \( b \) is the buffer power [dimensionless]; \( D \) is the diffusion coefficient in soil [cm\(^2\) s\(^{-1}\)] (\( = D_t f \), where \( D_t \) is the diffusion coefficient in free liquid and \( f \) is a tortuosity factor); \( s_o \) is the root radius [cm]; \( v \) is the effective velocity of flux solution [cm s\(^{-1}\)]; \( R_o \) is the initial half distance among root axis [cm]; \( J_m \) is the maximum influx [mol cm\(^{-2}\) s\(^{-1}\)]; \( K_m \) is the concentration for which the influx is \( J_m / 2 \) [mol cm\(^{-3}\)]; \( C_u \) is the threshold concentration below which influx stops [mol cm\(^{-3}\)]; \( R(t) \) is the half distance among roots axis [cm]; \( C_o(r) \) is the initial concentration profile in \([s_o, R_o]\) [mol cm\(^{-3}\)] and \( \ell(t) \) is the known root length as a function of time [cm] (the known law of root growth which can be linear, exponential or sigmoid; in the computed results we have used for some case the linear growth \( \ell(t) = \ell_o + kt \) with units of k in [cm s\(^{-1}\)] and the exponential growth defined by \( \ell(t) = \ell_o e^k \) with units of k in [s\(^{-1}\)]. We denote \( x_+ \) as the part positive of \( x \) defined by \( x_+ = \text{Max} (0;x) \). Equation (1) is the equation of diffusive and convective transport of ions in soil and condition (2) corresponds to the initial profile of
concentrations. Condition (4) represents a null flux on the moving limit of not-transference or instantaneous half mean distance between roots R(t). We remark that the null flux condition imposed in this paper by equation (4) is a more realistic condition and a corrected version with respect to the similar one used in our previous model [2]. Condition (3) represents the mass balance on the root surface and the expression (5) represents the moving boundary R(t) as a function of the instantaneous root length \( h(t) \). Expression (5) for the moving boundary is an improved version of a similar condition used in our previous model and introduces minor error in the computational algorithm designed to solve the problem. Unlike the expression proposed for R(t) in the previous version of our model which was based on considerations of constant volume of soil including roots (If the total volume of root plus soil remains constant then the amount of soil available to root will not be constant over time, i.e., the root grows at the expense of the decrease of the volume of soil), is now considered a constant volume of soil. The solution of problem (1) - (5) is obtained by the application of the adaptive finite element method (Schnepf et al. [19]) by using a dimensionless formulation through the following change of variable (similar to the one proposed by Roose, [10]), but now scaling the difference of coordinates \((r-s_0)\) by the difference of coordinates \((R(t)-s_0)\), i.e., transforming the variable interval \((s_0,R(t))\) in a fixed interval \((0,1)\) for all \( t > 0 \):

\[
C'(r^*, t^*) = \frac{C(r, t)}{K_m}, \quad r^* = \frac{r-s_0}{R(t)-s_0}, \quad t^* = \frac{D\phi}{(\phi + b)R_s^2} \quad t
\]

Once the influx values on the root surface are obtained we estimate the cumulative nutrient uptake by our growing root system by the following generalized formula valid for any range of concentrations (Reginato and Tarzia, [20]):

\[
\Delta U = U(t_i) - U(t_f) = 2\pi s_0 \int_{t_f}^{t_i} J(t) \ell(t) \, dt
\]

where \( J \) is given in mol cm\(^{-2} \) s\(^{-1} \), in cm and \( U \) in moles. The influx \( J(t) \) is given by:

\[
J(t) = J_m \left[ C(s_0,t) - C_u \right] / \left[ K_m + \left( C(s_0,t) - C_u \right) \right]
\]

Moreover, and based in this generalized formula; we define a weight averaged influx which is consistent with the experimental William’s formula [21]. This averaged influx is given by:

\[
\bar{J} = \frac{\int_{t_f}^{t_i} J(t) \ell(t) \, dt}{\int_{t_f}^{t_i} \ell(t) \, dt}
\]

where \( \bar{J} \) is given in mol cm\(^{-2} \) s\(^{-1} \). This weight averaged influx is more realistic because takes into account the temporal contribution of root length to the influx. In the case of constant influx \( J(t) = J \), the averaged influx given by (9) coincides with the temporal averaged influx.

3 THE SIMULATIONS

For comparison of simulations of influx on root surface and cumulative uptake versus observed data we use six set of input data sets extracted from literature. From now on, we denote the simulations as:

FB-NST 3.0: Original fixed boundary model NST 3.0,

First, we compute the influx on root surface of Cd by maize, sunflower, flax and spinach for two levels of concentration [3] and the results are shown in Table 1.

Table 1. Observed and predicted Cd influx by different crops to different soil Cd concentrations.

<table>
<thead>
<tr>
<th>Plant</th>
<th>Soil Conc. (10^-3 µmol cm^-3)</th>
<th>Influx of Cd (10^-16 mol cm^-2 s^-1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observed</td>
<td>Pred. NST 3.0</td>
</tr>
<tr>
<td>Maize</td>
<td>0.22</td>
<td>0.25</td>
</tr>
<tr>
<td>Sunflower</td>
<td>0.38</td>
<td>2.12</td>
</tr>
<tr>
<td>Flax</td>
<td>0.74</td>
<td>1.64</td>
</tr>
<tr>
<td>Spinach</td>
<td>0.48</td>
<td>7.55</td>
</tr>
<tr>
<td>Maize+</td>
<td>0.74</td>
<td>1.64</td>
</tr>
<tr>
<td>Sunflower+</td>
<td>1.80</td>
<td>5.56</td>
</tr>
<tr>
<td>Flax+</td>
<td>4.59</td>
<td>10.98</td>
</tr>
<tr>
<td>Spinach+</td>
<td>3.07</td>
<td>42.11</td>
</tr>
</tbody>
</table>

The influxes obtained by our MB-FE model are averaged by using the formula (9).

Second, we compute the influx on root surface and cumulative K uptake by maize, wheat, and sugar beet for low K soil and soil with K addition [13] and the results are shown in Tables 2 and 3.

Table 2. Observed and predicted cumulative K uptake for maize, wheat, and sugar beet grown on a low K soil with (+K) and without (-K) fertilization using data extracted from literature (Samal et al., 2010 [13])

<table>
<thead>
<tr>
<th>Cumulative Uptake (µmol)</th>
<th>Maize</th>
<th>Wheat</th>
<th>Sugar beet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FB-NST 3.0</td>
<td>380</td>
<td>1563</td>
<td>444</td>
</tr>
<tr>
<td>FB-NST 3.0*</td>
<td>1421</td>
<td>2410</td>
<td>443</td>
</tr>
<tr>
<td>MB-FE</td>
<td>1.56</td>
<td>867</td>
<td>23.4</td>
</tr>
<tr>
<td>R.L.D</td>
<td>0.9</td>
<td>0.95</td>
<td>0.95</td>
</tr>
</tbody>
</table>

Table 3. Observed and predicted root K influx on root surface for maize, wheat, and sugar beet grown on a low K soil with (+K) and without (-K) fertilization using data extracted from literature (Samal et al., 2010 [13])

<table>
<thead>
<tr>
<th>Influx on root surface (10^-12 µmol cm^-2 s^-1)</th>
<th>Maize</th>
<th>Wheat</th>
<th>Sugar beet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FB-NST 3.0</td>
<td>1.88</td>
<td>3.87</td>
<td>2.39</td>
</tr>
<tr>
<td>MB-FE</td>
<td>1.67</td>
<td>0.64</td>
<td>0.55</td>
</tr>
</tbody>
</table>
Third, we compute the influx on root surface and cumulative P uptake by peanut for low, intermediate and high soil concentrations [22] and the results are shown in Tables 4 and 5.

**Table 4.** Observed and predicted cumulative P uptake by peanut without root hairs at different soil levels using data extracted from literature (Singh et al., 2003 [22]).

<table>
<thead>
<tr>
<th>Cumulative Uptake (µmol)</th>
<th>Peanut</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 P</td>
</tr>
<tr>
<td>Observed</td>
<td>540</td>
</tr>
<tr>
<td>FB-NTS 3.0</td>
<td>1180</td>
</tr>
<tr>
<td>FB-NTS 3.0*</td>
<td>708.4</td>
</tr>
<tr>
<td>MB-FE</td>
<td><strong>468.8</strong></td>
</tr>
<tr>
<td>R.L.D. (72 days)</td>
<td>0.9→85</td>
</tr>
</tbody>
</table>

**Table 5.** Average predicted/observed ratio for influx of P by peanut on root surface and cumulative uptake obtained by NST 3.0, NST 3.0* and MB-FE using data extracted from literature (Singh et al., 2003 [22]).

<table>
<thead>
<tr>
<th>Influx on root surface (10^-8 µmol cm^-2 s^-1)</th>
<th>Peanut</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 P</td>
</tr>
<tr>
<td>Observed</td>
<td>3.44</td>
</tr>
<tr>
<td>FB-NTS 3.0</td>
<td>0.29</td>
</tr>
<tr>
<td>MB-FE</td>
<td><strong>1.15</strong></td>
</tr>
</tbody>
</table>

Finally, in order to verify the reliability of our moving boundary model and a 3D-dimensional architectural model we compute the P uptake by wheat to low concentrations [3] and the results are shown in Figures 1, 2 and 3.

![Figure 1](image-url) **Figure 1.** Observed, MB-FE and Heppel predicted values for the cumulative uptake of P by wheat seedlings over a 10 days period when grown in and low-P soils.
Figure 2. Predicted cumulative plant P acquisition by the MB-FE and the Heppel et al. model with an exponential branching distribution over a 90 days period when grown in and low-P soils.

Figure 3. Mass balance for the cumulative P uptake and the P ions remaining in soil with data of Heppel obtained by the moving boundary model.

4 RESULTS AND DISCUSSION

From Table 1, we conclude that MB-FE model predicts the average influx on root surface better than the FB-NST 3.0 always for different plants for two levels of concentrations. From Table 2, we conclude that for ions as K the MB-FE model is the best numerical method to compute the cumulative K uptake on a low K soil with K addition and without K addition for all cases. From Table 3, we conclude that in almost all cases, FB-NST 3.0 produces better predicted influxes except for maize without K addition. From Table 4 for low and increasing level of P concentration the fixed (original NST 3.0 and NST3.0*) and moving MB-FE models over predict always except in the soil without P addition (low concentrations). In this last case, MB-FE under predict with an acceptable error. From Table 5, we conclude that to low concentrations both models FB-NST 3.0 (with temporal average influx) and MB-FE (with weight average influx) under predict the average influx. For increasing level of P addition both models over predict the average influx. From Figure 1, we conclude that our MB-FE model predict better the final cumulative P uptake at 10 days. Although cumulative uptake predicted by the 3D model best fit to the experimental data we remark that this setting is obtained by choosing the best branching mode that fits to the experimental curve. Thus, the
The reason for which the MB-FE model is better than the other schemes (FB-NST 3.0 and FB-NST 3.0*) is that these last methods do not satisfy the mass balance among the ions taken by root and the ions remaining in soil. The ions remaining in soil were calculated by the following expression:

\[ N(t) = 2\pi(1+b)\int_{0}^{R(t)} rC(r,t)dr \]  

The mass balance for the program NTS 3.0 is not considered here because, obviously, to compute the ions remaining in soil, the operation must be done with concentration profiles as a function of time which has been calculated in fixed domain, but this result must be compared with the cumulative uptake by a growing root, i.e., which has been calculated by integration in a variable. Figure 3 show the mass balance for the results obtained by the finite elements method. The calculus was done with data extracted from [3] for the P uptake by wheat to low concentration. The obtained improvements by our model are mainly due to three factors: a) the use of a generalized formula for the cumulative nutrient uptake, b) the influxes obtained by the moving boundary model and the cumulative uptake, which are obtained through integration in a variable domain, while for the fixed boundary model the influxes are obtained in a fixed domain and the cumulative uptake by integration is on a variable domain, c) the use of a numerical method (finite element method) that ensures the balance of mass among the absorbed ions and the remaining ions in soil while the finite difference method does not satisfy it. Finally, in the light of these findings, conclusions drawn by previous papers [4] could be reinterpreted and our model could be included in larger field/catchment/climate scale models something which is not practically possible with the 3D numerical simulations due to their high computational burden.

REFERENCES


STUDY OF EFFECTS OF BLOOD AMINO ACID AND HORMONE LEVEL FOR CONTROLLING TRIGLYCERIDE ACCUMULATION IN THE LIVER OF RATS USING SELF-ORGANIZING MAP

M. MASUDA¹, Y. NAKABAYASHI², R. SHIOYA², H. NISHI³, S. TAKAHASHI³ AND F. HAKUNO³

¹The Center for Computational Mechanics Research, Toyo University
2100 Kujirai, Kawagoe, Saitama 350–8585, Japan
mst.masuda@gmail.com

²Faculty of Information Sciences and Arts, Toyo University
2100 Kujirai, Kawagoe, Saitama 350–8585, Japan
shioya@toyo.jp

³Graduate School of Agricultural and Life Sciences, The University of Tokyo
Bunkyo-ku, Tokyo 113-8657, Japan
ahakuno@mail.ecc.u-tokyo.ac.jp

Key words: Amino Acid, Triglyceride, Artificial Intelligence, Self-Organizing Map, Neural Network.

Abstract. Effects of blood amino acid and hormone level for triglyceride accumulation in the liver are revealed in some previous studies by experimental method. Since there are large individual differences in these effects, it needs a lot of rats for such experiments. Simulation can help, sometimes substitute experiment in various fields especially in engineering field. Recently, combining a supercomputer and Artificial Intelligence technique, simulation is expanding its scope. In this study, one of such a simulation technique, Self-Organizing Map (SOM), proposed by Kohonen and it is a kind of the Neural Networks and using for the competitive learning, is applied for classifying effects of blood amino acid and hormone level for controlling triglyceride accumulation in the liver of rats.

1 INTRODUCTION

Effects of blood amino acid and hormone level for triglyceride accumulation in the liver are revealed in some previous studies by experimental method³[¹][²][³]. Since there are large individual differences in these effects, it needs a lot of rats for such experiments. It takes time and cost a lot. In engineer field, simulations can help and substitute for experiments in various fields. We decided to identify experiments with simulations.

In this study, Simulations classify and distinguish by Self-Organizing Map (SOM), proposed by Kohonen[⁴][⁵]. SOM classifies effects of blood amino acid and hormone level to control Triglyceride accumulation in the liver. Decreases in experiments and cost can be expected. Classifications can be expected to give new findings such as change of blood amino acid increase or decrease Triglyceride in the liver.
2 EXPERIMENTAL DATA

We have already reported that protein malnutrition enhances insulin signaling in rat liver followed by accumulation of hepatic lipid\[^2\][\[^3\]. In addition, Aoyama et al have reported that triacylglycerol was accumulated in rat liver fed with low-arginine diet\[^1\]. From these backgrounds, we fed rats with various kinds of nutrition and measured level of blood amino acid, ammonia (NH\(_3\)), Ornithine (Orn), β-alanine (b-Ala), α-abscisic acid (α-ABA), Citrulline (Cit), taurine (Tau), Urea and triacylglycerol in liver (Liver TG). 5-week-old male Wistar male rats were purchased from Charles River Japan (Kanagawa, Japan). The rats were caged individually and kept in a room maintained at 24±1°C with 50-60% humidity and a 12h-light (8:00-20:00)/12h-dark (20:00-8:00) cycle. They were allowed free access to food and water throughout the experiment. All rats were fed control diet containing 15% (w/w) amino acids (detailed composition is shown in Table 1) for 4 days as a training and then divided into some experimental groups. For the next 7 days, each group was given either control diet or experimental diets in which total amino acids or only a specific amino acid was restricted to 1/3 of control diet (Table 1). In the experimental period, body weight and food intake of all rats were measured at 10:00 every day. In collecting the tissues, rats were anesthetized with isoflurane (DS Pharma Animal Health, Japan) and then blood and liver of each rat were obtained. Liver triglyceride was extracted with methanol:chloroform solution (1:2, v/v) according to the Folch’s method\[^6\] and quantified using Triglycerid E-test Wako (WAKO, Japan). Blood amino acid concentration was measured using Amino Acid Analyzer L-8900 (Hitachi Hi-Tech, Japan) according to the manufacturer’s protocol. All animal care and experiments conformed to the Guidelines for Animal Experiments of The University of Tokyo and were approved by the Animal Research Committee of The University of Tokyo.

**Table 1:** The composition of diets for rats

The composition of diet (g/kg) for each rat was shown. Soybean oil (50 g/kg), vitamin mixture (10 g/kg), mineral mixture (40 g/kg) and cellulose powder (100 g/kg) were added to each diet.

3 SOM

SOM, which is one of neural networks and was proposed by T. Kohonen, is a two-layered unsupervised competitive learning model that does not have a hidden layer. In the learning algorithm of SOM, the characteristics of input data are learned through neighborhood learning. A map is formed so that similar kinds of data are located in the neighborhood and other data
are located at distant places. Accordingly, it is possible to visualize high-dimensional vector data, and understand the relations among data intuitively.

As an example, the animals were classified. These characteristics of Table 2 are classified on the map (Fig.1).

### Table 2: Input Vectors of Animal

<table>
<thead>
<tr>
<th>Name</th>
<th>Small</th>
<th>Middle</th>
<th>Large</th>
<th>Nocturnality</th>
<th>2Legs</th>
<th>4Legs</th>
<th>Whiskers</th>
<th>Ungulate</th>
<th>Mane</th>
<th>Feather</th>
<th>Stripe</th>
<th>Hunt</th>
<th>Run</th>
<th>Fly</th>
<th>Swim</th>
<th>Herbivory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goose</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Fox</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Zebra</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Dove</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Wolf</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Goose</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Eagle</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Goose</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Dog</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Wolf</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Zebra</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Dove</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cat</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Owl</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Tiger</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.5</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Horse</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Hawk</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cow</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Figure 1:** Classified of Animals (16 kinds of animals with 16-dimensional characteristics.)

### 3.1 SOM Algorithm

SOM has a map with the M×N grid, and weighted units exist (Fig.2). Unit weights are updated in neighborhood learning. The algorithm of SOM is as follows:

**Step 1.** Initialize the weight vector w of a unit.

**Step 2.** Input the vector $\mathbf{x}_{\text{class}} = (x_{1\text{class}}, x_{2\text{class}}, x_{3\text{class}}, \ldots, x_{l\text{class}})$ into input layer. Here, “class” represents the number of input vectors, and “l” denotes the dimension of the input vector.

**Step 3.** Calculate the Euclidean distance “Dist” between an input vector and the weight vector of a unit in the map layer (competitive layer).
Step 4. Define the unit that minimizes the Euclidean distance obtained at Step 3 as the winning unit BMU (Best Matching Unit).

\[ BMU = \arg \min_m Dist_m \]  

Step 5. Update the weights of the winning unit and surrounding units only.

\[ \Delta w_m = h(l(BMU, m))(x^i_{class} - w_m^i) \]  

\[ h(l) = \alpha \exp \left( -\frac{l^2}{2\sigma^2} \right) \]  

Step 6. Repeat Steps 2 to 5 sufficiently.

Here, \( l \) represents the distance between the winning unit and the unit whose weight is updated. \( h \) and \( \sigma \) denote the neighborhood function and the range of the influence of the neighborhood function (neighborhood radius), respectively. This is decreased according to learning steps. \( \alpha \) is learning coefficient, ranging from 0 to 1. The learning coefficient \( \alpha \) decreases as learning steps proceed.

3.2 Input Data

Input Data given to SOM is created from the experimental data. Used data are blood amino acid in the rats and quantity of Triglyceride in the liver (Liver TG). Amino acid has 22 kinds, and the input data has 23 characteristics together with Triglyceride in the liver. Input of SOM give 23 inputs the combined blood amino acids and Liver TG per rat. Table 3 shows a part of input data. This can make it possible to estimate Liver TG by quantity of blood amino acid in a rat. Fig.3 shows the average values, the maximum values, and minimum values of the data.
Table 3: Blood amino acid and Liver TG of Rat

<table>
<thead>
<tr>
<th>Rat Number</th>
<th>Blood Amino Acid</th>
<th>Liver TG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Asp</td>
<td>Thr</td>
</tr>
<tr>
<td>No.1</td>
<td>0.119</td>
<td>5.693</td>
</tr>
<tr>
<td>No.2</td>
<td>0.122</td>
<td>7.698</td>
</tr>
<tr>
<td>No.3</td>
<td>0.143</td>
<td>6.058</td>
</tr>
<tr>
<td>No.4</td>
<td>0.113</td>
<td>5.014</td>
</tr>
<tr>
<td>No.60</td>
<td>0.098</td>
<td>6.48</td>
</tr>
<tr>
<td>No.61</td>
<td>0.075</td>
<td>3.815</td>
</tr>
<tr>
<td>No.62</td>
<td>0.088</td>
<td>6.248</td>
</tr>
<tr>
<td>No.63</td>
<td>0.053</td>
<td>5.424</td>
</tr>
</tbody>
</table>

Figure 3: Graph of Amino Acid and Liver TG (Dots are average values of each factor. Error bars respectively the maximum and minimum values.)

3.3 Learning SOM

In the SOM learning, map size was 30×30, and it was changed the neighborhood radius using RPSOM (Radius Parallel SOM). This method is similar to the temperature parallel SA. Several neighborhood radii are calculated in parallel, and it will select an appropriate neighborhood radius so as to satisfy the each evaluation value of Fitness and Interpolation. In this case, neighborhood radii were used three radii from 2.0 to 1.0, from 5.0 to 2.0 and from 8.0 to 3.0.

3.4 Results of SOM

Fig.4 show the learning result. The height direction of the figure represents the value of the Liver TG. At the center bottom of the map, there is mass of rats from No.22 to 27. At the center of left, cluster of no.40~45 and 50~55 is formed. These are judged these have similar characteristics. Table 3 indicates the quantity of blood amino acid and Liver TG of No.22, 25,
and 27 and No.40, 43, and 50. They are standardized at 0 to 1. Compared with No. 22, 25, and 27, Ser and Ala are respectively different. In addition, there are several items different from other two in each individual. It can be concluded they are very similar vectors on the whole. No.40, 43, and 50 were compared. Similarly, these also have the slightly different item respectively. There is individual difference, but it is a vector that is roughly similar. And Liver TG is also generally similar. From the Fig.4 (b), the classification map of Liver TG that has transferred smoothly was made.

![Figure 4](image-url)

**Figure 4:** Result of RPSOM (Map Size is 30×30, Input Vectors are 23, Learning Iterations are 1,000 times. (a) Color thickness indicates the distance from an input vector. The thin line denotes the distance between adjacent units. As it is nearer, the color whitens, and vice versa. (b) Height is Liver TG.

### Table 4: Blood amino acid and Liver TG of No22, 25, 27 and No.41, 43, 55

<table>
<thead>
<tr>
<th></th>
<th>Asp</th>
<th>Thr</th>
<th>Ser</th>
<th>Gly</th>
<th>Ala</th>
<th>Val</th>
<th>Met</th>
<th>Ile</th>
<th>Leu</th>
<th>Tyr</th>
<th>Phe</th>
<th>Lys</th>
</tr>
</thead>
<tbody>
<tr>
<td>No.22</td>
<td>0.104</td>
<td>0.924</td>
<td>0.896</td>
<td>0.852</td>
<td>0.714</td>
<td>0.482</td>
<td>0.044</td>
<td>0.407</td>
<td>0.420</td>
<td>0.404</td>
<td>0.651</td>
<td>0.317</td>
</tr>
<tr>
<td>No.25</td>
<td>0.115</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.844</td>
<td>0.525</td>
<td>0.041</td>
<td>0.547</td>
<td>0.535</td>
<td>0.443</td>
<td>0.725</td>
<td>0.387</td>
</tr>
<tr>
<td>No.27</td>
<td>0.103</td>
<td>0.916</td>
<td>0.743</td>
<td>0.823</td>
<td>0.590</td>
<td>0.469</td>
<td>0.022</td>
<td>0.501</td>
<td>0.450</td>
<td>0.370</td>
<td>0.731</td>
<td>0.280</td>
</tr>
<tr>
<td>No.40</td>
<td>0.160</td>
<td>0.254</td>
<td>0.503</td>
<td>0.154</td>
<td>0.018</td>
<td>0.693</td>
<td>0.692</td>
<td>0.084</td>
<td>0.518</td>
<td>0.183</td>
<td>0.177</td>
<td></td>
</tr>
<tr>
<td>No.43</td>
<td>0.159</td>
<td>0.232</td>
<td>0.525</td>
<td>0.190</td>
<td>0.025</td>
<td>0.653</td>
<td>0.672</td>
<td>0.088</td>
<td>0.497</td>
<td>0.208</td>
<td>0.253</td>
<td></td>
</tr>
<tr>
<td>No.50</td>
<td>0.123</td>
<td>0.294</td>
<td>0.152</td>
<td>0.023</td>
<td>0.655</td>
<td>0.605</td>
<td>0.076</td>
<td>0.364</td>
<td>0.194</td>
<td>0.231</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Asp</td>
<td>Thr</td>
<td>Ser</td>
<td>Gly</td>
<td>Ala</td>
<td>Val</td>
<td>Met</td>
<td>Ile</td>
<td>Leu</td>
<td>Tyr</td>
<td>Phe</td>
<td>Lys</td>
</tr>
<tr>
<td>No.40</td>
<td>0.950</td>
<td>0.086</td>
<td>0.378</td>
<td>0.671</td>
<td>0.469</td>
<td>0.476</td>
<td>0.035</td>
<td>0.744</td>
<td>0.567</td>
<td>0.711</td>
<td>0.782</td>
<td>0.456</td>
</tr>
<tr>
<td>No.43</td>
<td>0.835</td>
<td>0.127</td>
<td>0.395</td>
<td>0.543</td>
<td>0.478</td>
<td>0.506</td>
<td>0.029</td>
<td>0.793</td>
<td>0.603</td>
<td>0.654</td>
<td>0.857</td>
<td>0.524</td>
</tr>
<tr>
<td>No.50</td>
<td>0.915</td>
<td>0.156</td>
<td>0.542</td>
<td>0.609</td>
<td>0.537</td>
<td>0.668</td>
<td>0.045</td>
<td>0.880</td>
<td>0.649</td>
<td>0.591</td>
<td>0.812</td>
<td>0.111</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Asp</th>
<th>Thr</th>
<th>Ser</th>
<th>Gly</th>
<th>Ala</th>
<th>Val</th>
<th>Met</th>
<th>Ile</th>
<th>Leu</th>
<th>Tyr</th>
<th>Phe</th>
<th>Lys</th>
</tr>
</thead>
<tbody>
<tr>
<td>No.40</td>
<td>0.118</td>
<td>0.773</td>
<td>0.000</td>
<td>0.877</td>
<td>0.151</td>
<td>0.607</td>
<td>0.343</td>
<td>0.646</td>
<td>0.776</td>
<td>0.169</td>
<td>0.083</td>
<td></td>
</tr>
<tr>
<td>No.43</td>
<td>0.122</td>
<td>0.896</td>
<td>0.000</td>
<td>0.818</td>
<td>0.113</td>
<td>0.698</td>
<td>0.338</td>
<td>0.766</td>
<td>0.629</td>
<td>0.582</td>
<td>0.153</td>
<td></td>
</tr>
<tr>
<td>No.50</td>
<td>0.108</td>
<td>0.718</td>
<td>0.431</td>
<td>0.904</td>
<td>0.483</td>
<td>0.648</td>
<td>0.529</td>
<td>1.000</td>
<td>0.663</td>
<td>0.362</td>
<td>0.077</td>
<td></td>
</tr>
</tbody>
</table>
4 CONCLUSIONS

We tried to classify and estimate relation between the quantity of blood amino acid and Liver TG by using simulations, which used RPSOM. Obtained knowledge is indicated below.

- Classification of quantity of blood amino acid and Liver TG with RPSOM showed that the map arranged similar vectors was formed.
- As rat No.22 to 27 and No.40 to 45, 50 to 55, when quantity of blood amino acid are similar, Liver TG are also similar.
- The classification map of Liver TG that has transferred smoothly was made.
- It was possible to visually understand the relationship of quantity of blood amino acid and Liver TG using RPSOM. And that was classified into a similar group.

In the future, it will be necessary to consider predicting Liver TG from unknown input data and prediction method.

REFERENCES

ACCELERATED MULTI GPU TSUNAMI RUN-UP SIMULATION

SHOICHI FURUYAMA 1, YUSUKE MAIHARA 2, TSUKASA ASO 3

1 National Institute of Technology, Toyama College, Japan
Ebie-Neriya 1-2, Imizu, Toyama 933-0293, Japan
shoichi@nc-toyama.ac.jp, http://www.nc-toyama.ac.jp/WEB_Profile/shoichi/

2 National Institute of Technology, Toyama College, Japan
Ebie-Neriya 1-2, Imizu, Toyama 933-0293, Japan
yuusukesuke2002@yahoo.co.jp

3 National Institute of Technology, Toyama College, Japan
Ebie-Neriya 1-2, Imizu, Toyama 933-0293, Japan
aso@nc-toyama.ac.jp, http://www.nc-toyama.ac.jp/

Key words: GPU, TSUNAMI, Simulation, TESLA K20x, MPI, InfiniBand, Toyama-bay.

Abstract. Accelerated multi GPU Tsunami run-up simulation was realized at the wide area (whole Toyama-bay in Japan) by faster computation technique. To achieve the simulation, a high resolution calculation is required for the running-up Tsunami simulation because artificial structures on the ground such as roads, buildings, and houses are very small. On the other hand the huge area simulation is also required. In the Toyama-bay case the area is 42[km] x 15[km]. When 5[m] x 5[m] size computational cells are used for the simulation, over 26,000,000 computational cells are generated. To achieve the simulation, multi GPGPU system (NVIDIA TESLA K20x, InfiniBand, MVAPICH) was used for the calculation. As a result 5.16 times faster calculation was achieved on six GPUs than one GPU case and it was 86% parallel efficiency to the linear speed up.

1 INTRODUCTION

High resolution and high speed wave simulation was implemented on multi Graphics Processing Unit (GPU) system. The wave is assumed as Tsunami which is generated by earthquake at the seabed. Tsunami is one of the most dangerous phenomena for human life because it sometimes damaged at coastal region and runs-up on ground, moreover the speed of the wave is very fast.

The target area of the simulation is around Toyama-bay in Japan. The bay is located in middle of main island of Japan and faces the Sea of Japan. Almost of all area is over 1,000m in depth and it is one of the deepest bays in Japan and the bathymetry is suddenly dropped near the coast.

Tsunami has very long wave length and the speed is approximated by \( \sqrt{gh} \), here \( g \) is gravitational acceleration and \( h \) is depth. When the earthquake is occurred in the Toyama-bay, very fast Tsunami will hit the coastal area immediately and run-up on residential areas around
the bay. It is very serious scenario because of Toyama-bay has some active faults in the seabed.

A Tsunami simulation is one of the important technologies for protecting human life around coastal area. It predicts the Tsunami height at the coast and also provides the running-up area of the wave. Furuyama & Maihara simulated Tsunami wave at the part of Toyama-bay [1,2,3]. They got 78.5 times faster computation speed by using GPGPU than normal CPU calculation. However the simulation was not sufficient for the real phenomena because their target area was only around Toyama New Port and the initial wave was assumed as simple conditions. Therefore the purpose of this study is to extend the objective area and to use multi GPU system for the fast calculation of larger target area, i.e. whole Toyama-bay.

2 METHODOLOGY

2.2 GENERAL EQUATIONS

Tsunami wave is described by the following long-wave theorem, e.g., the continuity equation and the momentum equations.

\[
\frac{\partial \eta}{\partial t} + \frac{\partial M}{\partial x} + \frac{\partial N}{\partial y} = 0
\]

\[
\frac{\partial M}{\partial t} + \frac{\partial}{\partial x} \left[ \frac{M^2}{D} \right] + \frac{\partial}{\partial y} \left[ \frac{MN}{D} \right] + gD \frac{\partial \eta}{\partial x} + \frac{gn^2}{D^{7/3}} M \sqrt{M^2 + N^2} = 0
\]

\[
\frac{\partial N}{\partial t} + \frac{\partial}{\partial x} \left[ \frac{MN}{D} \right] + \frac{\partial}{\partial y} \left[ \frac{N^2}{D} \right] + gD \frac{\partial \eta}{\partial y} + \frac{gn^2}{D^{7/3}} N \sqrt{M^2 + N^2} = 0
\]

Here x and y are space coordination, \( h \) is static depth, \( g \) is gravity accelerate constant, \( \eta \) is elevation, \( D \) is total depth (=\( h + \eta \)) and \( n \) is Manning coefficient. The \( M \) and \( N \) are flux for x and y directions, respectively, and they are defined as follows.

\[
M = \int_{-h}^{\eta} u \, dz = u(h + \eta) = uD
\]

\[
N = \int_{-h}^{\eta} v \, dz = v(h + \eta) = vD
\]

At the ocean region, the non-linear long wave equation for momentum are approximated as linear long wave equation as follows.

\[
\frac{\partial M}{\partial t} + gD \frac{\partial \eta}{\partial x} = 0
\]
\[ \frac{\partial N}{\partial t} + gD \frac{\partial \eta}{\partial y} = 0 \]  

\[ \frac{\partial \eta}{\partial t} = \frac{1}{\Delta t} (\eta_{i,j}^{k+1} - \eta_{i,j}^k) \]  

\[ \frac{\partial M}{\partial x} = \frac{1}{\Delta x} \left( M_{i+1/2,j}^{k+1/2} - M_{i-1/2,j}^{k+1/2} \right) \]  

\[ \frac{\partial N}{\partial y} = \frac{1}{\Delta y} \left( N_{i,j+1/2}^{k+1/2} - N_{i,j-1/2}^{k+1/2} \right) \]  

The water level at next time step \((k+1)\) for linear long-wave equation are calculated as follow.

\[ \eta_{i,j}^{k+1} = \eta_{i,j}^k - \frac{\Delta t}{\Delta x} \left( M_{i+1/2,j}^{k+1/2} - M_{i-1/2,j}^{k+1/2} \right) - \frac{\Delta t}{\Delta y} \left( N_{i,j+1/2}^{k+1/2} - N_{i,j-1/2}^{k+1/2} \right) \]  

\[ M_{i+1/2,j}^{k+1/2} = M_{i+1/2,j}^{k-1/2} - gD_{i+1/2,j}^{k} \frac{\Delta t}{\Delta x} (\eta_{i+1,j}^k - \eta_{i,j}^k) \]  

\[ D_{i+1/2,j}^{k} = h_{i+1/2,j} + \frac{1}{2} (\eta_{i+1,j}^k + \eta_{i,j}^k) \]  

By using the similar process, the \(y\)-momentum equation and the half offset depth are discretized (Expression is omitted here).

In the case of the non-linear long-wave momentum equations are described as eq. (14). The upwind method is adopted for the momentum equations.
\[
M_{i+1/2, j}^{k+1} = M_{i+1/2, j}^{k} - \Delta t \left[ \lambda_{11} \left( \frac{M_{i+1/2, j}^{k-1/2}}{D_{i+1/2, j}^{k-1/2}} \right)^2 + \lambda_{21} \left( \frac{M_{i+1/2, j}^{k-1/2}}{D_{i+1/2, j}^{k-1/2}} \right)^2 + \lambda_{31} \left( \frac{M_{i+1/2, j}^{k-1/2}}{D_{i+1/2, j}^{k-1/2}} \right)^2 \right]
\]

Here,

\[
\begin{align*}
N_{i+1/2, j}^{k-1/2} &< 0, \quad \beta_{11} = 1, \quad \beta_{21} = -1, \quad \beta_{31} = 0 \\
\geq 0, \quad \beta_{11} = 0, \quad \beta_{21} = 1, \quad \beta_{31} = -1
\end{align*}
\]

\[
\begin{align*}
N_{i+1/2, j}^{k-1/2} &< 0, \quad \beta_{12} = 1, \quad \beta_{22} = -1, \quad \beta_{32} = 0 \\
\geq 0, \quad \beta_{12} = 0, \quad \beta_{22} = 1, \quad \beta_{32} = -1
\end{align*}
\]

\[
\begin{align*}
M_{i+1/2, j}^{k-1/2} &< 0, \quad \lambda_{11} = 1, \quad \lambda_{21} = -1, \quad \lambda_{31} = 0 \\
\geq 0, \quad \lambda_{11} = 0, \quad \lambda_{21} = 1, \quad \lambda_{31} = -1
\end{align*}
\]

\[
\begin{align*}
M_{i+1/2, j}^{k-1/2} &< 0, \quad \lambda_{12} = 1, \quad \lambda_{22} = -1, \quad \lambda_{32} = 0 \\
\geq 0, \quad \lambda_{12} = 0, \quad \lambda_{22} = 1, \quad \lambda_{32} = -1
\end{align*}
\]

Run-up condition is described by Iwasaki [4]. Overflowing condition is defined by Honma formula [5]. If the depth of both side of wall is defined as \( h_1, h_2 \) (\( h_1 > h_2 \)), the flux \( M \) is estimated as follows.

\[
M = \mu h_1 \sqrt{2gh_1} \quad \quad h_2 \leq \frac{2}{3} h_1
\]
\[ M = \mu' h_1 \sqrt{2g(h_1 - h_2)} \quad h_2 > \frac{2}{3} h_1 \]  

(20)

In the study, \( \mu = 0.36 \) and \( \mu' = 0.26 \mu \) are chosen.

2.3 GPU

TESLA K20x is used for the simulation as GPU. The system has two computational nodes and three GPUs are equipped in each node. Nodes are connected through InfiniBand (QDR, 40 Gb/s) like as figure 1. Table 1 shows the specifications of GPU and CPU.

![Figure 1: GPU Computing Server @NIT Toyama College (Left) & Overview of Computer System](image)

<table>
<thead>
<tr>
<th>Specification of GPU and CPU</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GPU</strong></td>
</tr>
<tr>
<td>NVIDIA TESLA K20X</td>
</tr>
<tr>
<td># of Cores</td>
</tr>
<tr>
<td>Frequency [GHz]</td>
</tr>
<tr>
<td>FLOPS (float) [TFLOPS]</td>
</tr>
<tr>
<td>Memory Bus Speed [GB/s]</td>
</tr>
</tbody>
</table>

2.4 GPGPU

Each computational cell is defined as the one-thread in the GPU. Because of a number of 26,223,750 computational cells are generated in the study (see Section 3.2), the number of threads is 26,223,750. CUDA (Driver Version 6.0, Runtime Version 5.5) was used for the developing environment.
3 TARGET

3.1 TARGET AREA

The target area of the simulation is Toyama-bay. Toyama-bay is located at middle of the main island of Japan and it faces the Sea of Japan (Figure 2). While the bay is half surrounded by the Noto peninsula, Tsushima warm current (TWC) intrudes from the Sea of Japan. Moreover the bay is one of the deepest bays in Japan and it has over 1,000 m in depth. The lower figure in Figure 2 shows the bathymetry of the bay. The deepest area is over 1,000 m in depth and the coastal region has steep seabed. The condition restricts to generate sand harbor. Coastal erosion is serious problem at the area. The buffer area is narrow between sea and residential area. So big waves sometimes hit the residential area around Toyama-bay. In the case of Tsunami, it is also serious problem. Because of the Tsunami energy is not decreased from the seabed friction, the wave hit the coastal area keeping huge energy. The estimation of tsunami waves by computer simulation around this area is important study for safety of resident.

![Figure 2: Simulation Area, Toyama-bay in Japan](image)

3.2 INITIAL CONDITIONS

Figure 3 shows the calculation area and the domain decomposition for the simulation. The size of calculation area is 41,625[m] x 15,750[m]. The size of computational cells are 5[m] x
5[m] and the total number of cells is 26,223,750. The red rectangular area shows the active fault named as Kurehayama active fault and the couple of percent possibility of earthquake is predicted within 30 years at the area. In the simulation the water level is increased to 5 [m] in 3 seconds as at the area as the initial condition. It means that the simple generating of tsunami wave for the simulation. The time step (dt) is set to 0.02 seconds and the total simulation time is set to 200 seconds, i.e., 10,000 steps are required for the simulation. The boundary conditions for ocean side are set as open boundaries and the boundary conditions for between sea and land are set run-up conditions ([4,5] and equations (19) and (20)). The domain decomposition is also shown in figure 3. The same size domains are distributed to the each six GPUs unit. The node1 (GPU1, 2 and 3) and the node2 (GPU4, 5 and 6) are connected by InfiniBand QDR and MPI is used for data transformation between nodes like as Figure 1.

![Figure 3: Domain Decomposition for Simulation Area](image)

4 RESULT

Figure 4 shows tsunami situation at 75 and 200 seconds after the earthquake is occurred. The high wave propagates circles (75 seconds) and hits the coastal region (200 second). The running-up waves (a couple of meters) are observed at some land regions where are residential areas in the simulation (200 second).

Table 2 shows the results of total calculation time, calculation time for each parts (water level and flux) , the message passing time and parallel efficiency with the some number of GPU devices. The simulation achieves 91.8 and 176.1 seconds calculation time by six and three GPUs which are the faster time than the real phenomena (200 seconds). The parallel efficiency means the ratio of the calculation time to the linear speed up. The high parallel efficiencies are achieved such as 90% and 86% to the linear speed up (ideal performance) in the cases of three and six GPUs parallel computing.
Table 2: Calculation Time and Parallel Efficiency to Linear Speed Up

<table>
<thead>
<tr>
<th># of GPUs</th>
<th>1</th>
<th>3</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Calculation Time [sec.]</td>
<td>473.3</td>
<td>176.1</td>
<td>91.8</td>
</tr>
<tr>
<td>Calc. for Water Level [sec.]</td>
<td>61.8</td>
<td>20.0</td>
<td>10.2</td>
</tr>
<tr>
<td>Calc. for Flux [sec.]</td>
<td>410.7</td>
<td>152.2</td>
<td>76.3</td>
</tr>
<tr>
<td>Message Passing Between Domains [sec.]</td>
<td>0</td>
<td>2.7</td>
<td>4.9</td>
</tr>
<tr>
<td>Parallel Efficiency (%)</td>
<td>N/A (100%)</td>
<td>90%</td>
<td>86%</td>
</tr>
</tbody>
</table>

5 DISCUSSION

The simulation result shows tsunami hits coastal area around Toyama-bay within 200 seconds after an earthquake occurs at Kurehayama active fault. The result indicates a very serious situation for residents around coastal area but the other hand the result provides useful information for making some evacuating scheme for residents.
The calculation time for the simulation is achieved in 91.8 seconds by using six GPUs. The parallel efficiency is high even in the case of six GPUs. The part of water level calculation, i.e. eq. (11), distributed well balanced load in each GPU. Figure 5 (Left) shows the water level calculation time at each time steps in each GPU. All GPUs' calculation time is around 1 [ms] and it shows the well balanced load distribution. As a result the speed up ratio is almost linear in the part. In the part of flux calculation, i.e. eq. (12) and (14) the calculation time at each time steps in each GPU is shown in Figure 5 (Right). There is 50% difference calculation time between GPU1 which has shortest calculation time and GPU2 which has longest one. The part has process of tsunami running-up calculation. Because of tsunami running-up areas (computational cells) are dynamically changing in the simulation, as a result the unbalanced load in each GPU influences the calculation performances. Moreover the flux calculation part occupies 83–87% of total calculation time and the load balancing affects directly the performance. However a high efficiency of parallel calculation was kept in the study. InfiniBand contributed the performance with its high speed data transfer ability between nodes and the data transfer time was only about 5 % of the total calculation time in the case of six GPUs were used. As the result the high performance of parallel efficiency was achieved. The result also indicated that the simulation time is shorter than the time of the real time phenomena (200 second). It will be able to apply the method to a real time (after earthquake occurring) tsunami disaster simulation in future.

![Figure 5: Load Balancing Situation. Calculation time of Flux (Left) and Water Level (Right). The horizontal axis means time steps and the vertical axis means calculation time of the part at each time step.](image)

6 CONCLUSIONS

Accelerated multi GPU Tsunami run-up simulation was realized at Toyama-bay in Japan with faster computation technique. The huge area simulation is required such as 42[km] x 15[km] area with fine computational cells of 5[m] x 5[m], as a result the simulation uses over 26,000,000 computational cells. To achieve the simulation, multi GPGPU system (NVIDIA TESLA K20x, InfiniBand, MVAPICH) was used for the calculation. The result was 5.16 times faster calculation on six GPUs than one GPU and it is 86% parallel efficiency to the linear speed up. The result also indicated that the simulation time is shorter than the time of the real time phenomena (200 second). It will be able to apply the method to a real time (after earthquake occurring) tsunami disaster simulation in future.
S. Furuyama, Y. Maihara and T. Aso

ACKNOWLEDGEMENT

This study has been supported by JSPS KAKENHI Grant-in Aid for Scientific Research C (Grant Number 24510260).

REFERENCES

VARYING THE BLOCK SIZE ON BLOCK CONJUGATE GRADIENT: COMPARISON OF STRATEGIES

PEDRO J. TORRES\textsuperscript{1}, CHRISTIAN E. SCHAERER\textsuperscript{2}, AMIT BHAYA\textsuperscript{3}

\textsuperscript{1}\textsuperscript{2}Polytechnic School National University of Asuncion, Campus de la UNA, San Lorenzo, Paraguay
P.O.Box: 2111 SL
torres.pedrozp@gmail.com, cschaer@pol.una.py

\textsuperscript{3} Federal University of Rio de Janeiro, Brazil.
Email: amit@nacad.ufrj.br

Key words: Block algorithms, adaptive, Block Conjugate Gradients, Ritz values.

Abstract. The Conjugate Gradient (CG) method has been widely used to solve linear systems of the type when the matrix $A$ is symmetric positive definitive. In the context of the implementation of CG as an iterative algorithm one important bottleneck is in the matrix vector multiplication which has the disadvantages of a low ratio of computation-memory accesses. To improve this ratio several strategies have been researched ranging from new paradigm machine architectures to optimizing the cache data usage through the usage of a matrix multivector multiplications Kernel. The latter in the CG context yields the Block Conjugate Gradient (BCG) originally used to solve linear systems with multiple right-hand-side. A main drawback of BCG for the case of a single right hand side is the determination of an optimal block size. Moreover, if an adequate block size is not chosen the effect on the rate of convergence of the CG can be negative and consequently its rate of convergence deteriorates. In this work we compare strategies for determining adaptively the optimal block size in the implementation of BCG considering a single right hand side seeking a trade-off between convergence rate and computational cost. The strategies analyzed are based on the Rayleigh Values associated to the search subspace, the associated Ritz Values, and the decay rate of the residual measured locally. Numerical experiments on real matrices are presented. It is observed that the Ritz Values strategy shows better results than the other strategies in terms of decreasing the computational cost. These results encourage further analysis and design of adaptive block algorithms.

1 INTRODUCTION

The solution of the large linear system of equations is the most time-consuming stage when solving many computational demanding industrial problems. Iterative methods are
widely used, constructed upon matrix-vector multiplication (MV) operations. Typically the cost associated with these operations is measured with two numbers: floating point operations and data communications (e.g. data transferred between memory and processor, or between processors).

Matrix-vector multiplications have very low ratio of flop over transferred bytes [14], specially when the matrix is sparse. This ratio is also called arithmetic intensity. On the other hand, modern computers have increased in processing speed (measured in flops) and have not been matched by the increase in speed of data communication. As a result, when performing a large number of MV multiplications we have idle processors waiting for data to be processed, wasting valuable computation time. Because of this, it is also said that MV operations are memory bound. For this reason, it is necessary to rethink the main kernel computation used in iterative algorithms, to exploit the computational resource available.

When seeking to improve the arithmetic intensity ratio, algorithms based on matrix-matrix multiplications kernel $Y \leftarrow AX + Y$ have good opportunities to increment this ratio, since the access to the matrix can be reduced, using for example blocking techniques at different levels [13, 14]. A very important case of this kernel is when the matrix $A$ is sparse and the matrices $X$ and $Y$ are dense with a small number of columns respect to the dimension of $A$. This is called Sparse Matrix-Matrix (SpMM) or Sparse Multivector Multiplications. Several investigations focused in optimizing the computational performance of this kernel has been conducted in recent years [2, 3, 1] achieving good results, even compared with the single vector case.

The use of this kind of kernel in numerical linear algebra was introduced in the early 80’s by Golub and Underwood in [11], inspired by eigenvalues problems with high clusterizations. O’Leary in [5] presented an extension of this block form to the conjugate gradients (CG) and related methods for solving simultaneously either multiple linear systems or a single linear system. Later, Nikishin et. al. in [7, 8] introduced a generalized block adaptive form for the conjugate gradient for solving a single linear system being possible to reduce the block size at each iteration. This algorithm is known as variable block CG (VBCG). However, experiments have demonstrated that in many cases, the reduction of the block size is not a good choice, particularly when new hardware trends are considered.

In this paper, we present a new alternative approach that adaptively chooses the block size in the VBCG. We also compare two different criteria: 1) simply increasing the block from a minimum size greater than one to a maximum value which depends on the hardware, and 2) an adaptive rule based on the Ritz Values which allows to increase the block size adaptively.

The remaining of this paper is organized as follows. In §2, we describe the block conjugate gradient (BCG) and the variable block conjugate gradient (VBCG), as well as it is introduced the Adaptive BCG based on Ritz values. A comparison between the methods for artificial matrices, as well as, for real matrices are presented in §3. Conclusions follow.
The numerical results show that the proposal criterion exhibit better performance in term of cost of SpMM, at least on the tested problem.

2 Adaptive Block Conjugate Gradient Algorithm

Block algorithms are based on the Block Krylov (BK) subspace, which is defined as follows

$$\mathcal{K}_{\text{Block}}(A, V, k) = \text{span}\{V, AV, \ldots, A^{k-1}V\},$$

where $V$ is a dense $n \times s$ matrix, such that $s \ll n$ and $\text{rank}(V) = s$. A Block Krylov subspace is a generalization of the well-known Krylov subspace (consider $s = 1$ in (1)).

The block CG method is an iterative process of minimization of a special non-negative functional

$$\min \text{tr}[(X - X^*)^T A (X - X^*)]$$

where $X, X^* \in \mathbb{R}^{n \times s}$ and $A \in \mathbb{R}^{n \times n}$, and $X^*$ is the global minimum of the unconstrained minimization problem. Equivalently, it is obtained the following matrix equation

$$AX - B = 0$$

where $B \in \mathbb{R}^{n \times s}$ and $X^*$ is the exact solution of the matrix equation.

The trivial approach is to solve (3) as $s$ separated and independent problems, using CG on each of them, i.e. solving $s$ problems $Ax_i = b_i$ with $x_i$ and $b_i$ being the columns of $X$ and $B$, respectively. This approach is equivalent of using the notion of line search method when solving (2). In this case, it is natural to considering the minimization of (2) over a larger subspace, using more available information, enriching the search space. This can potentially have a significant improvement in the resolution time. The block methods were developed with this philosophy. The BCG algorithm proposed by O’Leary in [5] for solving the matrix equation (3) is given by the Algorithm 1.

**Algorithm 1 Block Conjugate Gradient - BCG.**

1. Given an initial guess $X_0$ to the solution matrix $X^*$;
2. Set $R_0 = B - AX_0$, $P_0 = R_0\gamma_0$, where $\gamma_0 \in \mathbb{R}^{s \times s}$;
3. for $k = 0, 1, 2, \ldots$ do
   4. $X_{k+1} = X_k + P_k\alpha_k$
   5. $R_{k+1} = R_k - AP_k\alpha_k$
   6. $P_{k+1} = (R_{k+1} + P_k\beta_k)\gamma_{k+1}$
      where $\alpha_k$ and $\beta_k$ are computed as follows
   7. $\beta_k = (R_k^T R_k)^{-1}R_k^T R_{k+1}$
   8. $\alpha_k = (P_k^T AP_k)^{-1}R_k^T R_k$;
4. end for

In Algorithm 1, matrix $P_k$ is the search direction and $\alpha_k$ is the updating matrix parameter at iteration $k$. Matrix $R_k$ is the residual matrix of equation (3). An upper bound
for the convergence rate for the BCG algorithm is [5]:

$$e^{kT} A^k e_i \leq c \left( \frac{1 - \sqrt{\kappa}}{1 + \sqrt{\kappa}} \right)^{2k}, \quad 1 \leq i \leq s,$$

(4)

where $e_i^k = x^{k}_m - x^{k-1}_m$ is the $i$-th component of error at iteration $k$, $\kappa := \lambda_n/\lambda_s$ and $\lambda_n \geq \lambda_{n-1} \geq \cdots \geq \lambda_2 \geq \cdots \geq \lambda_1$ are the eigenvalues of the matrix $A$, and $c$ is a constant.

An interesting property of block method is observed in expression (4), since it implies that an increase of the block size $s$ potentially could lead to a faster convergence, reducing the number of accesses to the memory and decreasing the overall computational cost. However, an arbitrarily large increase of the block size could not compensate the arithmetic and communications costs, degrading the computational performance. Therefore a key point consists in choosing the block size $s$ to achieve a good trade off between the arithmetic intensity and the rate of converge of the algorithm.

In Algorithm 1 a static $s$ for all iterations can be chosen considering the worst case scenario to assure the convergence of the algorithm. The experience however shows that when using this strategy, many iterations need in fact a smaller $s$, wasting the computational time and degrading the arithmetic intensity. To exploit this situation, the variable block CG Algorithm 2 was proposed in [8, 7].

Algorithm 2 Variable Block Conjugate Gradient - VBCG.

1: Given an initial guess $X_0$ to the solution matrix $X^*$;
2: Set $R_0 = B - AX_0$, $P_0 = R_0 \gamma_0$, where $\gamma_0 \in R^{s \times s}$;
3: for $k = 0, 1, 2 \ldots$ do
4: $X_{k+1} = X_k + P_k \alpha_k$,
5: $R_{k+1} = R_k - AP_k \alpha_k$.
6: Choose with respect to some criterion a positive integer $s_{k+1} \leq s_k$.
7: Split the search subspace choosing a matrix $H_k$ such that, $\text{span}(P_k) = \text{span}(H_k) \oplus \text{span}(\tilde{P}_k)$, $P^T_k A H_k = 0$ and $H^T_k A H_k = I$.
8: Update the block direction $P_{k+1} = R_{k+1} + \tilde{P}_k \beta_k + \sum_{i=0}^{k} H_i \gamma_i^k$.
9: Update the matrices $R_{k+1}$ and $X_{k+1}$. $\alpha_k$, $\beta_k$ and $\gamma_k$ are computed such that
10: $R^T_{k+1} P_k = 0$, $P^T_{k+1} A \tilde{P}_k = 0$ and $P^T_{k+1} A H_i = 0$ for $i \leq k$, is satisfied respectively.
11: end for

The underlying strategy of the Algorithm 2 consists in removing a column of the matrix $P_k$ when a specific criterion is met, independently of rank the of the matrix, i.e. the search subspace is reduced by one dimension on each step and when a criterion rule is satisfied. The criterion established in [7] was to reduce the dimension of the matrix $P_k$ at every arbitrary constant number of iterations. Later in [8] a criterion based on eigenvalues of the matrix $P^T_k P_k$ was introduced.

In the context of this work we explore the connection between the convergence behavior of the CG (and BCG) and the distribution of the eigenvalues of the matrix $A$ for
designing a rule criterion for choosing $s$ adaptively. We claim that a criterion based on an approximation of the eigenvalues could be an intuitive choice, since the BCG intrinsically computes some approximation of eigenvalues at each iteration (Ritz Values). We are also interested in allowing the variation of $s$ within a range if this is necessary for maintaining the rate of convergence.

To achieve a more convergence-related approach in reducing or adding dimensions of the searching subspaces, we propose a new criterion for varying the block size. This criterion is based on the Ritz values implicitly computed by the CG algorithm. It is important to remark that the convergence behavior of CG methods is related to the quality of the Ritz values in terms of the approximation to eigenvalues of matrix $A$ [9, 10]. This fact was first pointed out in [10] and further demonstrated in [9].

This property was discovered when trying to explain the very fast convergence that CG exhibits in some problems. This is known as superlinear convergence. On the other hand, the Block version of Lanczos procedure accelerates the convergence of the Ritz values ([12], page 568) as well as being the preferred method when the matrix has multiple and clustered eigenvalues. Hence, we are led to think that an expansion or contraction of the search subspace could be determined as a function of the computed Ritz values on each iteration. In [6] was introduced a criterion for adjusting the block size in block algorithms as a function of the clustering of Ritz values when they are used for the eigenvalue problem. Our proposal in Algorithm 3 uses a similar criterion but in the context of the resolution of a single linear system using BCG.
Algorithm 3 Variable Block Lanczos Algorithm.

1. Let $A \in \mathbb{R}^{n \times n}$, a random $X_0 \in \mathbb{R}^{n \times p}$, $b \in \mathbb{R}^n$, block size $p \in \mathbb{R}$, and $tol \in \mathbb{R}^+$;
2. $B = b 1^T$, $1 \in \mathbb{R}^p$, $R_0 = B - AX_0$;
3. for $i = 0, 1, 2 \ldots$ do
   4. $M_k = X_k^TAX_k - X_{k-2}G_k^T$;
   5. $X_kG_k = R_k$;
   6. $Z = [Z \quad X_k]$;
   7. Compute $T = Z^TAZ$ and solve $T\psi = Z^Tb$
   8. Compute $X = Z\psi$ and $R = B - AX$
   9. if $\|r_j^{j+1}\| < \epsilon\|b\|$ for some $j \in [1, p]$ then break;
10. end if
11. if clustered then
12. Choose $\delta > 0$ and form an $n \times \delta$ matrix $U$ such that
13. $U^TX_{k+1} = 0$, $U^TX_i = 0$ and $U^TU = I_\delta$
14. $p_{j+1} = p_j + \delta_j$
15. $X_{k+1} = [X_{k+1}, U]$, $B_{j+1} = B_j + 1$
16. end if
17. else
18. $\delta = 0$;
19. end if
20. end for

Basically, the algorithm constructs an orthogonal basis and find a new solution solving the linear system projected onto this basis (steps 8 and 9). After this we evaluate the equation (5), if this is satisfied, e.g. clustered is true in step 13 of Algorithm 3, we expand the search subspace choosing one random vector and orthogonalizing against all the other vectors of the basis, as described in the following steps in Algorithm 3.

In order to measure the clusterization we use the following expression

$$\frac{\theta_{i+k-1} - \theta_i}{\min\{\theta_i - \theta_{i-1}, \theta_{i+k} - \theta_{i+k-1}\}} \leq \eta,$$

(5)

where $\theta_i$'s are the Ritz values at the iteration $j$, and $k$ is the size of the cluster being equal to the block size $p_j$. This condition is to test whether the Ritz values form a cluster of size $k$ based on the threshold $\eta$. An increment of the block size would lead to a better approximation of the Ritz values ([12] page 568). Then, based on [9], [10] improves the convergence rate of the BCG algorithm.

We also allow the Algorithm 3 to freely adjust the size of the block size, and in some cases the size is increased. The latter is justified by the observation in experimental results that in early iterations, a larger block size does not contribute to convergence.
3 EXPERIMENTAL RESULTS AND DISCUSSIONS

Experiment 1. We have chosen four matrices, with different and prescribed eigenvalue distributions. All the right-hand sides were chosen so that the initial residual has equal components in all eigenvector directions, so that no eigen-direction is favored. All the matrices tested in this experiment are diagonal matrices and an stopping tolerance $\|r\| < 10^{-9}\|b\|$. The matrices are the following:

- Matrix 1 (M1): is 100 × 100 matrix with eigenvalue distributions [0.0005 0.08 : 0.0115 : 1.21]. That is, one isolated eigenvalue 0.005 and the rest equally distributed in the interval [0.08, 1.21].

- Matrix 2 (M2): is 100 × 100 with eigenvalue distributions [1 : 0.5 : 50 : 400]. That is, one isolated eigenvalue 400 and the rest equally distributed in the interval [1, 50].

- Matrix 3 (M3): is 200 × 200 with eigenvalue distributions [1 2 400 : 1 : 597]. That is, two isolated eigenvalues 1, 2 and the rest equally distributed in the interval [400, 597].

- Matrix 4 (M4): is 904 × 904 with eigenvalue distributions [0.1 : 0.001 : 1 1.5 2.5]. That is, three isolated eigenvalues 1.5, 2, 2.5 and the rest equally distributed in the interval [0.1, 1].

This experiment was performed for fixed block size varying between $1 \leq p \leq 10$, and compared with proposal Algorithm 3. The results are shown in Figure 1. Observe at Figure 1, that there is a remarkable reduction in the number of iterations for matrices M1, M2 and M4; but for the matrix M3 which has its large eigenvalues clustered, the enrichment of the subspace has no effect. In addition, for all problems it is observed that large block size does not positively influence the rate of convergence at the beginning of the iterations process. For this reason, in the remain experiments we force the algorithm to start with small block size, and gradually increase it.

The increment in the average speed of convergence for matrices M1, M2, M3 and M5 with the block size could be understood as a consequence of the fact that the Block Lanczos algorithm finds more quickly clusters of eigenvalues, and that the CG algorithm is improved with the better approximations of the Ritz values. In other words, the bigger the block size is, the more able is the algorithm to approximate clusters of eigenvalues, leading to better convergence behavior of BCG. The computational cost and the loss of orthogonality in the process result in instability and impose a limit on the block size. For our proposal algorithm, the block size would be incremented automatically according to the criterion adopted, until a maximum block size of 8 is reached.

Experiment 2. Matrix from Real Applications.

The Table 1 summarize the matrices used for the experiment, extracted from University of Florida Sparse Matrix Collection[4].
Figure 1: Residual History for BCG and the Proposal Algorithms (ABL)

Table 1: Experiment 2. Matrix from Real Industry Applications

<table>
<thead>
<tr>
<th>Matrix</th>
<th>Application Area</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>bcsstk09</td>
<td>structural</td>
<td>1083</td>
</tr>
<tr>
<td>chem97ZtZ</td>
<td>statistical/mathematical</td>
<td>2541</td>
</tr>
<tr>
<td>poisson2500</td>
<td>computational fluid dynamics</td>
<td>2500</td>
</tr>
<tr>
<td>trefethen2000</td>
<td>combinatorial</td>
<td>2000</td>
</tr>
</tbody>
</table>

Figure 2 shows the results for algorithm 3 compared with the BCG algorithm. It is observed a significant improvement on the convergence rate in three matrix Trefethen2000, Chem97ZtZ, and bcsstk09. In all this matrices, we obtained a reduction of the number iterations increasing the block according to the criterion. This potentially could reduce the total computational cost, considering that the block increase correspond to a very low increase on the computational cost [2].

However, a remarkable lost of stability is observed in the iteration process for the same matrices. This could be due to the orthogonalization process, performed during adding an extra vector. This is not a minor issue, and must be considered in a further research.
4 CONCLUSIONS, ONGOING AND FUTURE WORK

In this paper, we have tested some strategies to vary the block size in the Block Conjugate Gradient algorithms. Some of them present encouraging results for the use of multivector optimization, like BCG and VBCG-ABL. While, the methods proposed in [7, 8] present instabilities due to rounding errors in some cases, we observed that the quality of the partition of the search space it is important to obtain convergence without increasing the computational complexity. In general, the variation and increasing block size approach used in Algorithm 3 presents good results, for isolated clusters. For distributed clusters of variable size along the real line the criteria have to be tuned. With respect to this point, the authors are working on a dynamic threshold to improve the performance of this kind of problems, as well as a combination of expansions and contractions of the search space to obtain a balance between the computational complexity and number of iterations.

REFERENCES


A DOMAIN-DECOMPOSITION BASED PARALLEL PROCEDURE FOR THE COMBINED FINITE-DISCRETE ELEMENT METHOD IN 2D

EARL E. KNIGHT¹, ZHOU LEI¹, ESTEBAN ROUGIER¹ AND ANTONIO MUNJIZA²

¹ Geophysics Group (EES-17), Los Alamos National Laboratory
P.O. Box 1663, 87545, New Mexico, United States
knighte@lanl.gov
zlei@lanl.gov
erougier@lanl.gov
http://www.lanl.gov

² Department of Engineering, Queen Mary, University of London
Mile End Road, London, E1 4NS, United Kingdom
a.munjiza@qmul.ac.uk - http://www.qmul.ac.uk

Key words: parallelization, FDEM, virtual engine, combined finite discrete element, fracture.

Abstract. Although the Combined Finite-Discrete Element Method (FDEM) has proven itself in dealing with problems of complex shapes, fracture and fragmentation, there is a stark reality of CPU requirements when dealing with industrial scale problems; in other words there is a compelling need for a parallel-processing framework to address large scale and grand challenge type of problems. One of the more recent development efforts in the context of FDEM was directed to implement the parallelization techniques needed for this method. In this paper a FDEM parallelization framework has been developed. Static domain decomposition and message passing inter-processor communication have been implemented in the FDEM code. The performance of the FDEM code in three typical problems is presented. For a discrete particle problem over 900 times speed-up has been obtained on 1000 processors. It has also been shown that the performance, especially efficiency of the parallelized software, still depends on the particular architecture of the computer hardware being used. However, the test results clearly demonstrate that it is possible to achieve very high efficiency even when a relatively large number of processors are employed.

1 2D FDEM AND PARALLELISATION

The Combined Finite-Discrete Element Method¹ (FDEM) merges the finite element based analysis of continua² with discrete element based transient dynamics, contact detection, and contact interaction solutions³. In FDEM the solid domains (called discrete elements) are discretized into finite elements, where finite rotations and finite displacements are assumed a priori and are formulated using a multiplicative decomposition based finite strain formulation. Through failure, fracture and fragmentation, single domains represented by separate finite element meshes are transformed into a number of interacting domains. The finite element discretization of solid domains is also conveniently used to discretize the contact between discrete elements.

In recent years, particularly in the field of rock mechanic applications, researchers have
employed the FDEM to simulate many complex industrial challenges\textsuperscript{[4]-[9]}. However, most of these works were accomplished in a single processor 2D realm. In order to tackle real world complex problems, a parallelization of the FDEM becomes mandatory. In this work, a domain decomposition-based parallel procedure for the FDEM in 2D is presented.

Domain decomposition is the most widely adopted approach to parallelization of Mechanics of Discontinua problems. The idea is simple: the physical space is divided into a fixed or moving semi-Eulerian grid of domain elements. Each domain element is assigned one processor. Physical objects, such as discrete elements are represented by objects of the object-oriented coding and are assigned to these domain elements according to their instantaneous (current) positions in space, Figure 1.

\textbf{Figure 1}: The concept of the domain decomposition approach – the physical space is divided into grid of computational domains.

With domain decomposition parallelization, only data at the boundaries of domain elements are exchanged or shared by different processors. “Speed ups” are limited by the number of elements allocated to a single processor. The larger the number of elements allocated to a processor, the longer the execution time from a CPU usage point of view. On the other hand, the larger the number of elements allocated to a single processor the smaller the relative communication (or data sharing) costs in terms of RAM usage. With more hardware available (CPUs), one obviously can achieve more speed.

That said, the challenge is to move from a hardware-oriented parallelization approach towards a problem-oriented parallelization approach. In this work, this goal is realized via the development and implementation of the Virtual Parallel Machine for FDEM (V-FDEM). The V-FDEM contains four domain decomposition engines: Virtual Space Engine, Virtual Boundary Engine, Virtual Gathering Engine, and Virtual Broadcasting Engine which are all that is needed to parallelize any domain decomposition FDEM job. The actual implementation of the process is made completely independent of any library such as MPI and is solely implemented using the C programming language.

\textbf{Virtual Space Engine}. The Virtual Space Engine is the simplest engine of V-FDEM. By
definition, it is composed of an existing sequential FDEM code. Details of how a sequential FDEM code is designed and implemented are outside the scope of this paper, but can be found in Munjiza\(^1\) or Munjiza et al.\(^3\).

**Virtual Boundary Engine.** The Virtual Boundary Engine handles and schedules data sending and receiving between processors. All the processors are therefore not talking (sending and receiving messages) in every communication time slot. However, there are no conflicts in communication between processors and all possible waiting is eliminated. This way, the data exchange time between processors becomes proportional to the number of discrete elements on the boundary of each domain; enabling good scaling performance.

**Virtual Gathering Engine.** As the data is assigned to different processors, there may be a need to gather all the data at a particular processor. This is done using the Virtual Gathering Engine. Within this framework the Virtual Gathering Engine commands each processor to merge the data that is already stored in the local memory of the processor with the data being sent by the sending processor.

**Virtual Broadcasting Engine.** This engine is used to disseminate information across the processors. When a single processor is broadcasting data, all data is passed to all the processors. However, each processor keeps only the part that it needs. Thus, data filtering is done at the receiving end. At the same time the broadcasting processor has a filter which decides what particular data from the database or input file needs to be broadcast.

## 2 NUMERICAL TESTS OF CPU PERFORMANCE

A representative sample of FDEM simulations has been run using up to 1000 processors and CPU performance has been measured. The first example is a borehole simulation without fracture. The second example is a borehole simulation with fracture. The third example is a large raster of deformable triangular bodies.

**Borehole Simulation without Fracture.** The performance of the parallel FDEM code was first tested on a borehole simulation. A sketch of the computational model used in this case, along with the time history of the pressure pulse applied to the borehole are shown in Figure 2. The physical domain consists of a 100 m size 2D square that has a 1 m diameter borehole in its center. The performance metrics for this problem are shown in Figure 3. It is worth noting that, in this case, the material is not allowed to fracture.
Earl E. Knight, Zhou Lei, Esteban Rougier and Antonio Munjiza

**Figure 2**: Borehole simulation: a) model setup; b) pressure time history.

**Figure 3**: Borehole simulation – CPU efficiency as a function of the number of processors.

**Borehole Simulation with Fracture.** The second example utilizes the same setup as the first one, but in this case the material is allowed to fracture. The performance metrics for this problem are shown in Figure 4. For illustration purposes, the fracture patterns at the end of the simulation are shown in Figure 5.
Earl E. Knight, Zhou Lei, Esteban Rougier and Antonio Munjiza

Figure 4: Borehole with fracture – CPU efficiency as a function of the number of processors.

Figure 5: Fracture patterns at the end of the simulation: a) general view; b) detail view.

Discrete Particles Raster. The third example consists of a $N \times N$ regular raster of triangular fully elastic particles interacting among them inside a rigid contained, as shown in Figure 6.

Figure 6: Discrete Particle Raster – Schematic view of a raster of $N \times N$ deformable equilateral triangles inside a square shaped rigid container.
Each triangle within the raster is provided with an initial random velocity, which changes from triangle to triangle, i.e., all three nodes of a given triangle have the same initial velocity prescribed to them. The problem has been run using different number of processors on a high performance computing Linux cluster. The performance metrics for this problem are shown in Figure 7.

![Figure 7](image.png)

**Figure 7**: Discrete Particle Raster – CPU efficiency as a function of the number of processors.

### 3 CONCLUSIONS

In this work it has been demonstrated that it is possible to build a problem-specific (as opposed to computer architecture-specific) virtual parallel machine for parallelization of existing FDEM software packages. Porting the parallel software to different parallel computer hardware architectures is thereby reduced to implementing some optimized aspects of the three remaining virtual engines and is in general, a relatively inexpensive operation.

The performance, especially efficiency of the parallelized software, still depends on the particular architecture of the computer hardware being used. However, the test results shown in this paper clearly demonstrate that it is possible to achieve very high efficiency even when a relatively large number of processors is employed.

### REFERENCES


APPLICATION OF DISCONTINUOUS GALERKIN ADAPTIVE MESH AND ORDER REFINEMENT METHOD TO ENERGY TRANSPORT AND CONSERVATION EQUATION IN RADIATION-HYDRODYNAMICS

M. HOLEC\textsuperscript{1,2}, M. COTELO\textsuperscript{3}, P. VELARDE\textsuperscript{3} AND R. LISKA\textsuperscript{2}

\textsuperscript{1}ELI-Beamlines Institute of Physics, AV CR, v.v.i
Na Slovance 2, Praha 8, 180 00, Czech Republic
milan.holec@eli-beams.eu

\textsuperscript{2}Czech Technical University in Prague, FNSPE
Brehova 7, 115 19 - Praha 1, Czech Republic
liska@siduri.fjfi.cvut.cz

\textsuperscript{3}Instituto de Fusion Nuclear, Universidad Politecnica de Madrid
Jose Gutierrez Abascal 2, Madrid 28006, Spain
manuel.cotelo@upm.es, pedro.velarde@upm.es

Key words: DG, FEM, AMR, non-local radiation transport, diffusion

Abstract. We focus on the energy conservation equation of the radiation-hydrodynamic model where the transport of radiation is present. Particularly, we aim to simulate space-time evolution of a coupled system of plasma temperature and radiation field in laser generated plasma. Our method is based on the finite element spatial discretization on rectangular meshes and ensures a precise higher-order calculation of radiation transport in transparent, non-local transport, and even in the diffusive plasma regimes, where the usual radiation transport codes based on the advection equation fail. We have further integrated an adaptive mesh refinement library, which implements parallel computation, into our code. Thus, we present a parallelized radiation transport 2D method which has the adaptive mesh and order refinement capability.

1 INTRODUCTION

In the field of laser generated high energy density physics, hydrodynamics is a standard approach to describe created plasma. We focus on the energy conservation equation of the radiation-hydrodynamic model [1]. Particularly, we aim to simulate space-time evolution of a coupled system of plasma temperature and radiation field which is induced by absorption of intense laser energy. As in [2, 3], we are challenging a specific behavior which exhibits the radiation transport, i.e. free-streaming photons propagation in...
transparent (optically thin) plasma regions and diffusion in high density (optically thick) plasma regions. Non-local transport occurs in between. Our work is based on an efficient adaptive mesh refinement (AMR) algorithm [4–7] where we add discontinuous Galerkin (DG) [8] order refinement capability. This extension is based on the high order spatial discretization library [9]. We want to ensure a precise calculation of radiation transport even in the diffusive plasma region, where the usual radiation transport codes based on the Boltzmann transport equation fail. This can be achieved with the help of adaptive mesh and order refinement (AMOR) method. Several numerical tests have been proposed to demonstrate code capabilities in literature. We mention two of very popular cases, i.e. the famous diffusive Marshak wave [10] and a non-equilibrium test of radiation transport equation [11]. In [12] there has been an extensive set of tests summarized and discussed. In this work we focus on a test presented in [3] which examines the code capabilities on the whole spectrum of photon transport conditions in plasma.

The paper is organized as follows: A high energy density physics appropriate continuum model of radiation-hydrodynamics is given in Sec. 2. In Sec. 3 we present a model of radiation transport and energy conservation equation. Special attention is given to the asymptotic solution of radiation transport equation. Sec. 4 deals with the DG numerical scheme of radiation transport and energy conservation equation and its properties are discussed. In Sec. 5 we refer to the adaptive mesh refinement method and its extension by the DG numerical scheme. In order to demonstrate the capabilities of our proposed scheme we compare the results with an analytical case, which is addressed in Sec. 6. We conclude with the summary of our results in Sec. 7.

2 RADIATION-HYDRODYNAMICS MODEL

In radiation-hydrodynamics, we aim to describe our physical system by using two different structures, i.e. the plasma fluid and the radiation. These are coupled by emission and absorption of photons. Defined in Lagrangian frame such a system governs a set of five equations [1]

\[
\frac{d\rho}{dt} + \rho \nabla \cdot \mathbf{v} = 0, \tag{1}
\]

\[
\rho \frac{d\mathbf{v}}{dt} + \frac{1}{c^2} \frac{d\mathbf{F}}{dt} = -\nabla p - \nabla \cdot \mathbf{P}, \tag{2}
\]

\[
\rho \frac{d\epsilon}{dt} + p \nabla \cdot \mathbf{v} + \frac{du}{dt} + \nabla \cdot \mathbf{F} = -\nabla \cdot \mathbf{L} - \nabla \cdot \mathbf{q}, \tag{3}
\]

where \(\rho\) is the fluid density, \(\mathbf{v}\) the fluid velocity, \(p\) the fluid pressure, \(\epsilon\) the fluid internal energy density per mass, \(\mathbf{L}\) the energy flux given by laser, \(\mathbf{q}\) the heat flux, \(c\) the speed of light, \(u\) the radiation energy density, \(\mathbf{F}\) the radiation flux, and \(\mathbf{P}\) the radiation pressure tensor. In order to close the radiation-hydrodynamic system, one needs to know \(p = p(\epsilon, \rho)\) and \(\mathbf{P} = \mathbf{P}(u)\), i.e. the equation of state (EOS) and the radiation pressure closures.
3 RADIATION TRANSPORT AND ENERGY CONSERVATION

In order to relate the radiation field quantities in (3, 2) to radiation transport we use the intensity of photons $I_{\nu}^{\prime} = I(t, x, n, \nu)$, which is a scalar function describing the quantity of photons traveling in the direction $n$ and having energy given by the frequency $\nu$. We write down the definition of radiation energy density, flux, and pressure tensor depending on $I_{\nu}^{\prime}$, i.e.

$$u = \frac{1}{c} \int_{\nu} \int_{4\pi} I_{\nu}^{\prime} \, d\mathbf{n} \, d\nu,$$

(4)

$$F = \int_{\nu} \int_{4\pi} n \, I_{\nu}^{\prime} \, d\mathbf{n} \, d\nu,$$

(5)

$$P = \frac{1}{c} \int_{\nu} \int_{4\pi} n n \, I_{\nu}^{\prime} \, d\mathbf{n} \, d\nu,$$

(6)

which helps us relate the transport of photons to the radiation-hydrodynamic model (2, 3).

3.1 Radiation transport equation

Equations (3) and (2) contain radiation field quantities (4), (5), and (6) depending on radiation intensity $I_{\nu}^{\prime}$. We adopt the Boltzmann transport equation of photons interacting with plasma [2] to calculate $I_{\nu}^{\prime}$, i.e.

$$\frac{1}{c} \frac{\partial I_{\nu}^{\prime}}{\partial t} + \mathbf{n} \cdot \nabla I_{\nu}^{\prime} = k_{\nu}(f_{\nu}(T) - I_{\nu}^{\prime}),$$

(7)

$$I_{\nu}^{\prime}(\mathbf{x}) = I_{\nu}^{\prime}\mathbf{n}_{\Gamma}(\mathbf{x}), \forall \mathbf{x} \in \Gamma^-,$$

(8)

where $k_{\nu}$ is absorption coefficient (inverse of the $\nu$-energetic photon mean free path), $f_{\nu}(T)$ is the source of photons emitted from plasma, and the boundary condition applies at $\Gamma^-$, which is the segment of boundary where the intensity $I_{\nu}^{\prime}$ enters, i.e. $\mathbf{n} \cdot \mathbf{n}_{\Gamma} < 0$. The functions $f_{\nu}$ and $k_{\nu}$ are frequency dependent, and furthermore, we expect the source function to be temperature dependent and generally anisotropic.

3.2 Energy conservation equation

The energy transport model may be simplified by gathering the radiation non-relevant energy sources together. This means we put together the divergence of laser and heat flux and we also add the hydrodynamic work term to obtain the source $S = -\nabla \cdot \mathbf{L} - \nabla \cdot \mathbf{q} - p \nabla \cdot \mathbf{v}$. We also define the energy dependence on temperature $\rho \frac{\partial \varepsilon}{\partial T} = a \frac{\partial T}{\partial t}$. Using the latter assumptions and with the help of (4, 5) we can write the energy conservation equation (3) as temperature dependent in the form

$$a \frac{\partial T}{\partial t} + \int_{\nu} \int_{n} \left( \frac{1}{c} \frac{\partial I_{\nu}^{\prime}}{\partial t} + \mathbf{n} \cdot \nabla I_{\nu}^{\prime} \right) \, d\mathbf{n} \, d\nu = S,$$

(9)
where the radiation intensity $I_{n}^{\nu}$ is given by the transport equation (7). One can observe that $I_{n}^{\nu}$ also depends on temperature, which is given by the source function $f_{n}^{\nu}(T)$ in (7).

### 3.3 Asymptotic behavior of energy transport model

Asymptotic behavior means that the solution of (7) depends on the absorption coefficient $k_{\nu}$, more precisely, the solution tends to be advective when $k_{\nu} \to 0$, i.e. the transparent plasma, and tends to the diffusion when $k_{\nu} \to \infty$, i.e. the diffusive plasma. The first case leads to pure advection of intensity $I_{n}^{\nu}$ for any direction $n$ and for any frequency $\nu$, i.e.

$$\frac{1}{c} \frac{\partial I_{n}^{\nu}}{\partial t} + n \cdot \nabla I_{n}^{\nu} = 0 .$$

(10)

This directly implies the energy equation (9) to be

$$a \frac{\partial T}{\partial t} = S .$$

(11)

This means that radiation and plasma evolve as two independent systems. The second case has been analyzed in [2]. The transport equation (7) can be solved as a power series in the small parameter $1/k_{\nu}$. This leads to the asymptotic solution

$$I_{n}^{\nu}(x) = f_{n}^{\nu}(T) - \frac{1}{k_{\nu}} n \cdot \nabla f_{n}^{\nu}(T) + O(k_{\nu}^{-2}) .$$

(12)

When we put (12) into (9) and assume $f_{n}^{\nu}(T)$ to be isotropic, we obtain after some algebra

$$a \frac{\partial T}{\partial t} + 4 \pi c \int_{\nu} \frac{\partial f_{n}^{\nu}}{\partial T} \frac{\partial T}{\partial t} - \nabla \cdot \int_{\nu} \frac{4 \pi}{3 k_{\nu}} \frac{\partial f_{n}^{\nu}}{\partial T} \frac{\partial T}{\partial t} \frac{\partial T}{\partial t} = S ,$$

(13)

which is the famous Rosseland radiation diffusion limit [2] modeled by heat equation with a nonlinear diffusion coefficient given by $\int_{\nu} \frac{4 \pi}{3 k_{\nu}} \frac{\partial f_{n}^{\nu}}{\partial T} \frac{\partial T}{\partial t} d\nu$.

### 4 DISCONTINUOUS GALERKIN DISCRETIZATION

#### 4.1 Radiation transport equation discretization

In order to discretize (7) we use the strong form discontinuous Galerkin variational principle on a functional space $H$

$$\int_{\Omega^{\nu}} \left( \frac{1}{c} \frac{\partial I_{n}^{\nu}}{\partial t} + n \cdot \nabla I_{n}^{\nu} + k_{\nu} I_{n}^{\nu} - k_{\nu} f_{n}^{\nu} \right) \psi \, d\Omega = \int_{\Gamma^{\nu}} (F_{n} - F_{n}^{*}) \cdot n \psi \, d\Gamma , \forall \psi \in H ,$$

(14)

where we have applied the Green’s formula twice. As the boundary condition, we use the difference between the natural flux and the numerical flux, i.e. $(F_{n} - F_{n}^{*})$. In our particular case of radiation intensity, we know that the natural flux has the following form

$$F_{n} = n I_{n}^{\nu} .$$

(15)
Since we are treating the transport of photons traveling with speed of light, we need to use the upwind numerical flux in order to make our method to be numerically stable for the time step corresponding to hydrodynamics. That is, we choose

\[
F^*_{n} = \begin{cases} 
    nI^\nu_n, & n \cdot n_\Gamma > 0 \\
    n\tilde{I}^\nu_n, & n \cdot n_\Gamma < 0 
\end{cases},
\]  

(16)

which means that the outgoing flux depends on the internal intensity \(I^\nu_n\) and the incoming flux is given by some external intensity \(\tilde{I}^\nu_n\).

The spatial discretization applies to the functional space \(H\). We shrink the space to a finite basis

\[
\psi = [\psi_1, \ldots, \psi_N]^T
\]

(17)

to a discrete space \(H_N^\psi\). In \(H_N^\psi\), we can find the following representation of the functions

\[
I^\nu_n = [I^\nu_n, \ldots, I^\nu_N]^T,
\]

(18)

\[
f^\nu_n = [f^\nu_n, \ldots, f^\nu_N]^T,
\]

(19)

and their spatial evaluation

\[
I^\nu_n(x) = \psi(x)^T \cdot I^\nu_n,
\]

(20)

\[
f^\nu_n(x) = \phi(x)^T \cdot f^\nu_n,
\]

(21)

where the source function \(f^\nu_n(x)\) can be interpolated with an arbitrary function set

\[
\phi = [\phi_1, \ldots, \phi_N]^T.
\]

(22)

Consequently, we have a discrete system for the intensity \(I^\nu_n\) calculation

\[
\int_{\Omega^\nu} \psi \otimes \left( \psi^T \frac{1}{c} \frac{\partial I^\nu_n}{\partial t} + n \cdot \nabla \psi^T \cdot I^\nu_n + k^\nu \psi^T \cdot I_n - k^\nu \phi^T \cdot f^\nu_n \right) \mathrm{d}\Omega = \int_{n \cdot n_\Gamma < 0} \psi \otimes \left( \psi^T \cdot I^\nu_n - \tilde{\psi}^T \cdot \tilde{I}^\nu_n \right) n \cdot n_\Gamma \mathrm{d}\Gamma. 
\]

(23)

That is, we can calculate \(I^\nu_n\) for any direction \(n\) and any frequency \(\nu\).

The scheme may be written using matrices and backward Euler method as

\[
\frac{1}{c} M^\psi \cdot \frac{I^\nu_{n+1} - I^\nu_n}{\Delta t} + n \cdot G^\psi \cdot I^\nu_{n+1} + k^\nu M^\psi \cdot I^\nu_{n+1} - k^\nu M^\phi \cdot f^\nu_{n+1} = \\
\Gamma^\psi_n \cdot I^\nu_{n+1} - \tilde{\Gamma}^\psi_n \cdot \tilde{I}^\nu_n,
\]

(24)

\[\text{Example Image: Figure 1 (not shown)}\]
where the matrices used are the following

\[ M^{\psi\psi} = \int_{\Omega} \psi(x) \otimes \psi^T(x) \, d\Omega, \quad (25) \]

\[ G^{\psi\psi}_t = \int_{\Omega} \psi(x) \otimes \frac{\partial \psi^T(x)}{\partial x_i} \, d\Omega, \quad (26) \]

\[ \Gamma^{\psi\psi}_n = \int_{n \cdot n_\Gamma < 0} \psi(x) \otimes \psi^T(x) \, n \cdot n_\Gamma \, d\Gamma, \quad (27) \]

\[ \tilde{\Gamma}^{\psi\psi}_n = \int_{n \cdot n_\Gamma < 0} \psi(x) \otimes \tilde{\psi}^T(x) \, n \cdot n_\Gamma \, d\Gamma, \quad (28) \]

\[ M^{\psi\phi} = \int_{\Omega} \psi(x) \otimes \phi^T(x) \, d\Omega. \quad (29) \]

Finally we show the discrete solution of (7)

\[ \left( \left( \frac{1}{c \Delta t} + k_\nu \right) M^{\psi\psi} + n \cdot G^{\psi\psi} - \Gamma^{\psi\psi}_n \right) \cdot I^{\nu n+1} = k_\nu M^{\psi\phi} \cdot f^{\nu n+1} + \frac{1}{c \Delta t} M^{\psi\psi} \cdot I^{\nu n} - \tilde{\Gamma}^{\psi\psi}_n \cdot \tilde{I}^{\nu n}. \quad (30) \]

At this moment we want to express the explicit dependence of the source function \( f^{\nu n}(T) \) on temperature. In a discrete form we can write

\[ k_\nu f^{\nu n+1} = B^{\nu n} \cdot T^{n+1} + h^{\nu n}, \quad (31) \]

where the matrix \( B^{\nu n} \) and the vector \( h^{\nu n} \) depend on the direction \( n \) and the frequency \( \nu \). We are particularly interested in the explicit dependence of intensity on temperature. This dependence can be expressed from the solution of radiation transport equation (30) where we make use of (31). Finally, we obtain a linear dependence of \( I^{\nu n+1} \) on temperature \( T^{n+1} \), i.e.

\[ I^{\nu n+1} = A^{\nu n} \cdot T^{n+1} + b^{\nu n}, \quad (32) \]

where the explicit formulas of the solution coefficients are

\[ L = \left( \left( \frac{1}{c \Delta t} + k_\nu \right) M^{\psi\psi} + n \cdot G^{\psi\psi} - \Gamma^{\psi\psi}_n \right), \quad (33) \]

\[ A^{\nu n} = L^{-1} \cdot M^{\psi\phi} \cdot B^{\nu n}, \quad (34) \]

\[ b^{\nu n} = L^{-1} \cdot \left( \frac{1}{c \Delta t} M^{\psi\psi} \cdot I^{\nu n} - \tilde{\Gamma}^{\psi\psi}_n \cdot \tilde{I}^{\nu n} + M^{\psi\phi} \cdot h^{\nu n} \right). \quad (35) \]

One should notice that we construct and invert the transport operator (33) on every computational cell separately.
4.2 Energy equation discretization

We apply the Galerkin variational principle on energy conservation equation (9), i.e.

$$\int_{\Omega} \phi \left( a \frac{\partial T}{\partial t} + \int_{\nu} \int_{\mathbf{n}} \left( \frac{1}{c} \frac{\partial I_{\nu}^{\mathbf{n}}}{\partial t} + \mathbf{n} \cdot \nabla I_{\nu}^{\mathbf{n}} \right) d\mathbf{n} d\nu - S \right) d\Omega = 0, \forall \phi \in H.$$  

The process similar to (17) delivers a temperature spatial discretization

$$\int_{\Omega} \phi \phi^{T} \left( \frac{a}{\Delta t} \frac{\partial T}{\partial t} + \int_{\nu} \int_{\mathbf{n}} \psi^{T} \cdot \frac{1}{c} \frac{\partial I_{\nu}^{\mathbf{n}}}{\partial t} + \mathbf{n} \cdot \nabla \psi^{T} \cdot I_{\nu}^{\mathbf{n}} d\mathbf{n} d\nu - \phi^{T} \cdot S \right) d\Omega = 0 (36)$$

where the discrete function $T$ alongside the source function $f_{\nu}^{\mathbf{n}}$ belongs to the finite functional space $H_{N}^{\phi}$ given by the basis (22). Application of the backward Euler method and numerical integration over frequencies $\nu$ and directions $\mathbf{n}$ as discrete ordinates gives us the scheme

$$M^{\phi \phi} \cdot \frac{T_{n+1}^{\nu}}{\Delta t} + \sum_{\nu, \mathbf{n}} \frac{1}{c} M^{\phi \psi} \cdot \frac{I_{\nu}^{\mathbf{n+1}} - I_{\nu}^{\mathbf{n}}}{\Delta t} + \mathbf{n} \cdot G^{\phi \psi} \cdot I_{\nu}^{\mathbf{n+1}} \Delta \mathbf{n} \Delta \nu = M^{\phi \phi} \cdot S, \quad (37)$$

where we use the following matrices

$$M^{\phi \phi} = \int_{\Omega} \phi(x) \otimes \phi^{T}(x) d\Omega, \quad (38)$$
$$G^{\phi \psi}_{i} = \int_{\Omega} \phi(x) \otimes \frac{\partial \psi^{T}}{\partial x_{i}}(x) d\Omega, \quad (39)$$
$$M^{\phi \psi} = \int_{\Omega} \phi(x) \otimes \psi^{T}(x) d\Omega. \quad (40)$$

We can further develop the scheme by using (32), and eventually, we obtain the temperature $T_{n+1}$ variable scheme

$$\left( \frac{a}{\Delta t} M^{\phi \phi} + \sum_{\nu, \mathbf{n}} \frac{1}{c} M^{\phi \psi} \cdot A_{\nu}^{\mathbf{n}} + \mathbf{n} \cdot G^{\phi \psi} \cdot A_{\nu}^{\mathbf{n}} \Delta \mathbf{n} \Delta \nu \right) \cdot T_{n+1} = \frac{a}{\Delta t} M^{\phi \phi} \cdot T_{n} + M^{\phi \phi} \cdot S + \sum_{\nu, \mathbf{n}} \frac{1}{c} M^{\phi \psi} \cdot I_{\nu}^{\mathbf{n}} + \left( \frac{1}{\epsilon \Delta t} M^{\phi \phi} + n \cdot G^{\phi \psi} \right) \cdot b_{\nu}^{\mathbf{n}} \Delta \mathbf{n} \Delta \nu. \quad (41)$$

Let us summarize some interesting properties of the discretization scheme (41) which solves the energy equation (9). It has no explicit dependence on $I_{\nu}^{\mathbf{n}}$. Indeed, time evolution of radiation intensities for all directions and frequencies is inherently included in the scheme due to (32). That is, by evolving the temperature in time we also implicitly evolve the radiation field. This is done for every cell separately while the inter-cell communication is done by means of $I_{\nu}^{\mathbf{n}}$ in (30). The quantity $I_{\nu}^{\mathbf{n}}$ must be then iterated over the whole computational mesh to align the natural and the numerical fluxes in (14). This makes the scheme conservative in energy.
Figure 1: An AMR grid structure example. Transport of photons in a transparent plasma spatially discretized by using three mesh spacing levels. These different mesh spacing levels are successively overlaid one on another. From the left: colormap on the most refined level, then colormap on the union of two most refined levels and finally colormap on the union of all three levels. The inner boundary condition between levels is passed by making use of the matrix (42).

5 ADAPTIVE MESH AND ORDER REFINEMENT

We solve the equations (7) and (9) in 2D, for a sequence of embedded uniform Eulerian meshes of different mesh spacings. Each mesh spacing is represented by a level of gathered boxes, i.e. meshes. Such levels are then hierarchically assembled to create an AMR grid structure. This grid structure is generated by BoxLib [4] where it aims to keep the numerical error constant by focusing on steep gradients in temperature and radiation intensity. An example of such a AMR grid structure is shown in Figure 1, where an assembly of three levels consisting of boxes (meshes) can be seen. In order to work with the AMR scheme, information has to be transferred among meshes, and also, among different mesh spacing levels. This information transfer is nothing else than inner boundary condition, i.e. the numerical flux used in (14) in case of radiation transport equation solved by the scheme (24). Nevertheless, one must be aware of a more general case of information transfer between two levels. Indeed, we are challenged by hanging nodes, which appear due to different mesh spacings. In order to solve the latter problem we must generalize the boundary matrix (28) in the following manner

$$\tilde{\Gamma}_n^\psi \tilde{\phi}_i = \sum_{\Delta \Gamma_i} \int_{\mathbf{n} \cdot \mathbf{n}_{\Gamma_i} < 0} \psi(x) \otimes \tilde{\phi}_i^T(x) \mathbf{n} \cdot \mathbf{n}_{\Gamma_i} \, d\Gamma_i,\quad (42)$$

where $\Delta \Gamma_i$ is a mutual part of boundary between the proper element identified by the basis $\psi$ and an outer element identified by the basis $\tilde{\phi}_i$. Furthermore, we directly obtain the order refinement capability of our method based on (42), where we apply a higher-order basis $\psi$ compared to the basis $\tilde{\phi}_i$.  

926
Figure 2: Steady state solution to (43) calculated using quadrilateral elements Q1 and Q3. Photons are generated by the source function \(\sin(\pi y)\) and travel in the direction from left to right. We show two different values of \(k \in \{1, 1e4\}\) which correspond to non-local and diffusive radiation transport, respectively. One can observe, that in the diffusive case, i.e. \(k = 1e4\), the intensity of photons aligns with the source function \(\sin(\pi y)\), which is in agreement with (12). Table 1 shows quantitative errors and convergence order of our method.

6 NUMERICAL TEST

The slab geometry implies that the source function and the boundary condition do not depend on \(x\) coordinates. We define \(\mu = \cos(\theta)\), where \(\theta\) is the angle between the photon propagation direction \(\mathbf{n}\) and \(y\)-axis. We apply the zero boundary condition, i.e. \(I(y_0, \nu) = 0, \forall \nu\), and the steady state, i.e. \(\frac{\partial I}{\partial t} = 0\). Consequently, we obtain a simplified form of (7), an ordinary differential equation [3]

\[
\mu \frac{dI(y, \mu)}{dy} = k (f(y) - I(y, \mu)) \quad (43)
\]

We can find a formal solution to (43) as

\[
I(y, \mu, y_0) = \int_{y_0}^{y} f(\tilde{y}) \exp \left( \frac{k}{\mu} (\tilde{y} - y) \right) \frac{k}{\mu} d\tilde{y} \quad (44)
\]

In our particular case, we will use the source function

\[
f(y) = \sin(\pi y) \quad (45)
\]
defined on the interval \( y \in (0, 1) \). The corresponding solution of (44) reads

\[
I(y, \mu, y_0) = \exp\left(\frac{k (y_0 - y)}{\mu}\right) \left(\frac{k \pi \mu \cos(\pi y_0)}{\mu^2 \pi^2 + k^2} - \frac{k^2 \sin(\pi y_0)}{\mu^2 \pi^2 + k^2}\right) + \frac{k^2}{\mu^2 \pi^2 + k^2}\left(\sin(\pi y) - \frac{\mu \pi}{k} \cos(\pi y)\right).
\]

(46)

Table 1 shows the errors and convergence order of our method without AMR. The results correspond to intensity \( I(y, \cos(\pi/4), 0) \) compared with the analytic solution (46). It is shown that our method is high-order accurate for values \( k \in (1e^{-4}, 1e^4) \) covering transparent, non-local and diffusive transport regime in plasma, see Figure 2. We use the \( L_1 \) norm on a rectangular mesh to analyze our scheme.

### Table 1: Steady state test errors and convergence order. Our method proves to be high-order accurate for values \( k \in (1e^{-4}, 1e^4) \) covering transparent, non-local and diffusive transport regimes in plasma.

<table>
<thead>
<tr>
<th>element</th>
<th>cells</th>
<th>( E_{L1}^{k=1e^{-4}} )</th>
<th>( q_{L1}^{k=1e^{-4}} )</th>
<th>( E_{L1}^{k=1e^0} )</th>
<th>( q_{L1}^{k=1e^0} )</th>
<th>( E_{L1}^{k=1e^4} )</th>
<th>( q_{L1}^{k=1e^4} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>10</td>
<td>2.7e-07</td>
<td></td>
<td>2.3e-03</td>
<td></td>
<td>8.3e-03</td>
<td></td>
</tr>
<tr>
<td>Q1</td>
<td>20</td>
<td>4.9e-08</td>
<td>2.5</td>
<td>4.3e-04</td>
<td>2.4</td>
<td>1.6e-03</td>
<td>2.4</td>
</tr>
<tr>
<td>Q1</td>
<td>40</td>
<td>1.2e-08</td>
<td>2.0</td>
<td>1.1e-04</td>
<td>2.0</td>
<td>3.7e-04</td>
<td>2.1</td>
</tr>
<tr>
<td>Q1</td>
<td>80</td>
<td>2.9e-09</td>
<td>2.0</td>
<td>2.6e-05</td>
<td>2.0</td>
<td>9.0e-05</td>
<td>2.1</td>
</tr>
<tr>
<td>Q2</td>
<td>10</td>
<td>4.6e-09</td>
<td></td>
<td>4.1e-05</td>
<td></td>
<td>3.5e-07</td>
<td></td>
</tr>
<tr>
<td>Q2</td>
<td>20</td>
<td>4.1e-10</td>
<td>3.5</td>
<td>3.5e-06</td>
<td>3.5</td>
<td>5.2e-08</td>
<td>2.8</td>
</tr>
<tr>
<td>Q2</td>
<td>40</td>
<td>4.5e-11</td>
<td>3.2</td>
<td>4.0e-07</td>
<td>3.2</td>
<td>1.2e-08</td>
<td>2.1</td>
</tr>
<tr>
<td>Q2</td>
<td>80</td>
<td>5.4e-12</td>
<td>3.1</td>
<td>4.7e-08</td>
<td>3.1</td>
<td>2.8e-09</td>
<td>2.1</td>
</tr>
<tr>
<td>Q3</td>
<td>10</td>
<td>7.3e-11</td>
<td></td>
<td>2.6e-07</td>
<td></td>
<td>2.3e-06</td>
<td></td>
</tr>
<tr>
<td>Q3</td>
<td>20</td>
<td>2.8e-12</td>
<td>4.7</td>
<td>8.4e-09</td>
<td>5.0</td>
<td>1.0e-07</td>
<td>4.5</td>
</tr>
<tr>
<td>Q3</td>
<td>40</td>
<td>1.5e-13</td>
<td>4.2</td>
<td>4.3e-10</td>
<td>4.3</td>
<td>5.6e-09</td>
<td>4.2</td>
</tr>
<tr>
<td>Q3</td>
<td>80</td>
<td>8.9e-15</td>
<td>4.1</td>
<td>2.4e-11</td>
<td>4.1</td>
<td>3.3e-10</td>
<td>4.1</td>
</tr>
</tbody>
</table>

We are particularly interested in the diffusive condition, i.e. when \( k \gg 1 \). In that case the first term of (46) behaves as \( \exp(-y/\mu k) \). The second term goes in limit to \( \sin(\pi y) - \frac{\mu \pi}{k} \cos(\pi y) \), which is nothing else than the asymptotic formula \( I \approx f - \mu/k \frac{df}{dy} \) which corresponds to (12). In order to discuss the numerical results corresponding to diffusion we consider \( k = 1e^4 \). Then we evaluate the anisotropic part of the second term of (46)

\[
-\frac{k^2}{\mu^2 \pi^2 + k^2} \frac{\mu \pi}{k} \cos(\pi y) \approx 1e^{-4},
\]

(47)

where \( k = 1e^4, \mu = \cos(\pi/4) \).
With this estimate we can discuss the importance of the order of our method in the case of diffusive regime. The second-order element $Q_1$ is not efficient, because it gives an error $9.0e-05$ on 80 cells spatial resolution, which is of the same order as the diffusive term (47), while the error $2.3e-06$ on 10 cells spatial resolution delivered by the fourth-order element $Q_3$ is in the order of percents of (47), i.e. $Q_3$ satisfies the diffusion regime even on such a low spatial resolution.

7 CONCLUSIONS

We have developed a 2D adaptive mesh and order refinement method on rectangular meshes. It satisfies all proposed radiation transport conditions, i.e. it follows both asymptotic solutions of the radiation transport equation and it is conservative. The method is stable, and furthermore, it is computationally efficient thanks to the AMR capability and the parallelization of BoxLib. We have also analyzed the importance of order refinement, which plays a crucial role in the diffusion regime of radiation transport in plasma.

REFERENCES


MULTI-FRONTAL MULTI-THREAD DIRECT SOLVER 
WITH GALOIS SYSTEM FOR ADAPTIVE FINITE 
ELEMENT METHOD

Anna Paszyńska¹, Konrad Jopek², Maciej Woźniak³, Maciej Paszyński⁴, 
Donald Nguyen⁵, Andrew Lenerth⁶, Keshav Pingali⁷

¹Jagiellonian University 
Krakow, Poland 
anna.paszynska@uj.edu.pl

²AGH University of Science and Technology 
Krakow, Poland 
kjopek@gmail.com

³AGH University of Science and Technology 
Krakow, Poland 
macwozni@agh.edu.pl

⁴AGH University of Science and Technology 
Krakow, Poland 
maciej.paszynski@agh.edu.pl

⁵Institute for Computational Engineering and Science, The University of Texas 
Austin, USA 
ddn@cs.utexas.edu

⁶Institute for Computational Engineering and Science, The University of Texas 
Austin, USA 
lenharth@ices.utexas.edu

⁷Institute for Computational Engineering and Science, The University of Texas 
Austin, USA 
pingali@cs.utexas.edu

Key words: Adaptive finite element method, multi-frontal direct solver, graph grammar, GALOIS, tree rotation

Abstract. In this paper we present our multi-frontal multi-thread graph grammar based direct solver with GALOIS scheduler for two dimensional adaptive grids with point and edge singularities. We utilize efficient elimination trees constructed for the adaptive grids, and we tests their efficiency by utilizing the tree rotation algorithm and comparing the quality of the elimination trees before and after the rotation.
1 Introduction

In this paper we focus on application of graph grammar based multi-thread multi-frontal direct solver [8] with GALOIS scheduler [10] for an efficient solution of two-dimensional adaptive grids [5]. We utilize on the state-of-the-art multi-frontal direct solver [6, 7]. The problem of construction of the optimal elimination tree controlling the execution of the solver is NP-complete [11]. In this paper we test the quality of the elimination trees obtained from dynamic programming algorithm [1] as well as reproduced by heuristic algorithm [9] utilized for an efficient solution of two dimensional grids with with point and edge singularity [5]. We also test our graph grammar based solver [8, 9] by comparison with state-of-the-art MUMPS solver [2–4].

2 Tree rotation algorithm.

We also propose the following algorithm improving the balancing of the elimination tree. The algorithm browses the elimination tree in BFS and performs a sequence of local rotations each time one of the tree branches is much deeper than the other one. The routine should be called $\text{Rotate}(T, \text{NULL})$ where $T$ is the root of the tree. The algorithm utilizes the topological data obtained from the computational mesh. The routine returns the difference of the depths of the right and left the branches of the tree.

```c
1 int Rotate(root, parent)
2 {
3     if left branch is present then root->l = rotate(left) else root->l = 0
4     if right branch is present then root->r = rotate(right) else root->r = 0
5     h=root->r - root->l
6     if (h == 1 || h == -1 || h == 0) then return max(root->l,root->r)+1
7     T=NULL
8     if (h >= 2) then
9         //rotations to the left
10         child = child of root->right, neighbor of root->brother, having smaller number of neighbors
11         other_child = other child of root->right
12         if (child==root->right->right) && (root->right->r<=root->right->l) ||
13             (child==root->right->left) && (root->right->l<=root->right->r) then
14             //rotation to the left
15             T = rotate_left(root, parent, child, other_child)
16         }
17         else
18         {
19             parent1 = root
20             root1 = root->right
21             child1 = child of root1->left, neighbor of root1->brother,
```

The algorithm uses the following subroutines:

```c
int recalculate_tree_depth(root)
{
    if left branch is present
        then root->l = recalculate_tree_depth(left)
        else root->l = 0
    if right branch is present
        then root->r = recalculate_tree_depth(right)
        else root->r = 0
    return max(root->l,root->r)+1
}
```

```c
node rotate_left(root, parent, child, other_child)
{
    T = root->right
    t = root->left
    if T==NULL return root
    root->right = t
    root->left = child
    T->right = other_child
    return T
}
```

The algorithm is illustrated on the meshes with point and edge singularities, obtained from the execution of the heuristic algorithm. Let us focus first on the tree for the mesh with point singularity, presented in Figure 1.

1. The algorithm browses the elimination tree in BFS and checks the balancing of nodes.

2. It finally finds a node that has left branch depth equal to 4 and right branch depth equal to 2.
3. It performs rotation to the right. It rotates elements 7, 8 since they have same number of common edges with 11 as 9, 10.

4. Next, it finds a node with left branch depth equal to 3 and right branch depth equal to 5.

5. It performs double rotation to the left. It rotates 5, 6 towards 11 since they have smaller number of common edges.

6. Next, it finds a node with left branch depth equal to 5 and right branch depth equal to 2.

7. It performs rotation to the right. It rotates 5, 6, 11 towards 12 since they have common edges.

8. Next, it finds a node with left branch depth equal to 3 and right branch depth equal to 6.

9. It performs double rotation to the left. It rotates 12 towards 3, 4 since they have one common edge.

10. Next, it finds a node with left branch depth equal to 5 and right branch depth equal to 2.

11. It performs rotation to the right. It rotates 3, 4, 12 towards 13 since they have common edges.

12. Next, it finds a node with left branch depth equal to 3 and right branch depth equal to 6.

13. It performs rotation to the left. It rotates 3, 4, 12, 13 towards 1, 2 since they have common edges.

Let us focus first on the tree for the mesh with edge singularity, presented in Figure 2.

1. The algorithm browses the elimination tree in BFS and checks the balancing of nodes.

2. It finally finds a node that has left branch depth equal to 6 and right branch depth equal to 4.

3. It performs rotation to the right. It rotates elements 13, 14, 15, 16, 19, 20, 21, 22 since they have same number of common edges with 5, 6, 17, 18 as 1, 2, 3, 4, 7, 8, 9, 10.

4. Next, it finds a node with left branch depth equal to 6 and right branch depth equal to 3.
5. It performs double rotation to the left.

Let us compare the elimination trees before and after the rotations. The Figure 3 presents the elimination tree obtained by the heuristic algorithm, and its version after the execution of the tree rotation. The unrotated tree is unbalanced, in the sense that left branch depth is equal to 2, while the right branch depth is equal to 8. The rotated tree on the contrary is well balanced, its left branch has depth 5 and the right branch has branch 4. Additionally, Figure 4 presents the elimination tree obtained by our heuristic algorithm, and its version after the execution of the tree rotation algorithm. Again, the unrotated tree is unbalanced, its left branch depth is equal to 7, while the right branch depth is equal to 3. The rotated tree is well balanced, its left branch has depth 5 and the right branch has branch 4.
3 Numerical results

3.1 Comparison with MUMPS

In this example we compare our GALOIS solver with heuristic trees without the rotation algorithm against MUMPS solver with AMD ordering provided by METIS, since AMD results in minimum number of FLOPs. We have compiled our GALOIS based solver with gcc-4.8.0 compiler. Our solver does not use any optimized numerical libraries, it is a pure C code. We compare against MUMPS solver version 4.10.0 compiled with gfortran-4.8.0, linked with metis-4.0.3, atlas-3.10.1, LAPACK-3.4.2, ScaLAPACK-2.0.2. In the MUMPS solver we utilize Cholesky factorization (the problem is setup as symmetric positive definite). We use simple model heat transfer equation. We utilize AMD ordering algorithm with MUMPS, since it provides fastest execution times according to our measurements. In our solver we utilize LU factorization. We compare execution times as well as parallel efficiency and speedup. The tests are performed on a single node of ATARI linux cluster with 16 core Intel(R) Xeon(R) CPU, with 2.4 GHz, total 16 GB RAM.

The comparison of the execution times for point singularity is presented in Figure 5. Let us summarize the performed numerical experiment:

- The number of degrees of freedom and the execution times for 2D mesh with point singularity is small, thus FLOPs is not a dominating factor, rather initializations, memory transfers, cache usage etc.
**Figure 5**: Comparison of the execution times of the MUMPS and GALOIS solvers for different numbers of threads, for different numbers of refinement levels, for the mesh with point singularity. Comparison of speedup of both solvers.

- Both MUMPS and GALOIS solvers slow down when we increase number of threads for such the small problem.

- The fastest is one thread MUMPS. However, one thread GALOIS is as fast as two threads MUMPS, and generally four and eight threads MUMPS solvers are slower than any threads GALOIS solvers.

In the following experiments we focus on the mesh with edge singularity. The comparison of the execution times for edge singularity and the parallel efficiency of the solvers is presented in Figure 6.

We can draw the following conclusions from the performed numerical experiments:

- For small problem sizes (less than 10000 degrees of freedom) the MUMPS and GALOIS solvers behave like for point singularity case.

- For larger problem sizes both MUMPS and GALOIS speeds up when we increase number of cores.

- For larger problem sizes GALOIS scales well up to 8 cores, however MUMPS scales well up to 4 cores only.
For larger problems the GALOIS solver with any thread number outperforms MUMPS multi-thread solvers.

### 3.2 Tests of the tree rotation algorithm

In this section we present comparison of the execution times for the meshes with point and edge singularities, using our heuristic elimination trees, before and after execution of the rotation algorithm. The tests are performed on a GILBERT machine with 64 cores. The comparison is presented in Figures 7 and 8. We can make the following conclusions from the performed numerical experiments:

- The execution time for the solver with elimination tree after the rotation is sometimes faster than the execution time of the solver with the elimination tree before the rotation. But sometimes it is also slower. Actually the execution times of the two versions of the solver oscilates.

- This is true for both point and edge singularity.

- There is no big difference in the execution times between the trees before and after rotation, and thus we conclude that our heuristic tree are well suited for parallel processing.
Figure 7: Comparison of the execution times of the GALOIS solver with heuristic elimination trees before and after execution of the rotation algorithm for different numbers of threads, for different numbers of refinement levels, for the mesh with point singularity.

Figure 8: Comparison of the execution times of the GALOIS solver with heuristic elimination trees before and after execution of the rotation algorithm for different numbers of threads, for different numbers of refinement levels, for the mesh with edge singularity.
4 Conclusions

In this paper we described a graph grammar based implementation of the multi-frontal solver for adaptive finite element method with GALOIS scheduler. We tested the graph grammar based solver on two exemplary computational problem, two dimensional mesh with point singularity as well as two dimensional mesh with edge singularity. The elimination trees has been obtained from the heuristic algorithm [9] and verified by the dynamic programming exhaustive search [1]. However, the heuristic algorithm as well as dynamic programming approach considered optimal elimination trees for sequential execution of the solver. The trees that are optimal for the sequential execution of the solver, may be not optimal for the parallel execution. This is related to well balancing of the elimination trees. In this paper we propose a tree rotation algorithm, that browses the elimination trees in BFS order, and improves iteratively the balancing of the elimination tree. The resulting elimination tree are indeed more balanced in the sense that number of tree nodes on left and right sub-branches are almost identical at all layers of the tree. We compared the execution times of the graph grammar based solver working on the elimination trees before and after the rotation algorithm. We conclude that the rotated trees results in similar behavior of the parallel solver. This is because indeed the trees are more balanced in terms of the number of tree nodes, however the computational costs in rotated trees may be not uniformly distributed. Even with current version of the tree rotation algorithm and the elimination trees, we outperform sequential and parallel state-of-the-art MUMPS solver. The future work may involve development of better tree rotation algorithms, taking into account not only the number of nodes, but also the number of resulting floating-point operations in both branches of the tree. However, such the algorithm is hard to design, since it’s computational cost may be larger than the cost of solution itself.

5 Acknowledgements

The work presented in this paper is supported by Polish National Science Center grant no. DEC-2012/07/B/ST6/01229. The work of Keshav Pingali and Andrew Lenerth was supported by NSF grants CCF 1337281, CCF 1218568, ACI 1216701, and CNS 1064956. Donald Nguyen was supported by a DOE Sandia Fellowship.

References


[2] Amestoy P.R. and Duff, I.S. Multifrontal parallel distributed symmetric and unsym-


USE OF MAXIMUM ENTROPY SHAPE FUNCTION WITHIN A STRONG MESHFREE FORMULATION.

FELIPE A. MARCHANT.¹ AND FRANCO O. PERAZZO.²

¹Aula UTFSM-CIMNE, Universidad Técnica Federico Santa María
Avenida España #1680, Valparaíso - Chile
felipe.marchant@usm.cl

²Universidad Técnica Federico Santa María, Departamento de Ingeniería Mecánica
Avenida España #1680, Valparaíso - Chile
franco.perazzo@usm.cl

Key words: MAXENT, shape function, strong formulation, meshfree, collocation

Abstract. We present a new approach for the strong formulation methods, using for that the maximum entropy (maxent) meshfree approximants as a shape functions that are positive, smooth and with uniform aspect ratio even for nonuniform node sets. We study the performance of this type of shape function in a collocation scheme to obtain the numerical solution of a number of PDE’s, including elliptic second-order PDE and one diffusion case. The examples highlight advantage of this type of shape function and shows that can be competitive when they are compared with the traditional used in collocation scheme.

1 INTRODUCTION

In the past years, in the context of the meshfree method (MF), it has been established a clear predominance of the methods based on a Weak or Galerkin Formulation (WF) over the methods based on a strong formulation (SF) [1, 2, 3]. This has been mainly motivated for the fact that the WF has resulted in a more stable methodology, and the only problem present in these methods are the scheme and domain of integration [4, 5]. In the case of SF the situation is completely different, because they have had to deal with the problem of instability and not robustness. One of the major factors responsible for this, is the differential operator which is characterized as an error amplifier. Another drawback of such methods is the difficulty of imposed the Neumann Boundary Condition or Derivative condition, which involve another equations group different that those obtained for the domain. This problem has been reported for many authors [6, 7, 8, 2, 3]. Among the solutions proposed to deal with these issues of instabilities on SF, the main line of research is how the local
subdomains for the approximation are built, which has led to positive results in some particular case [9, 10]. We present in this work a different type of solution consistent in the use of a different shape function instead of the traditional weight least squares (WLS) used on SF. To obtain this shape function the Maximum Entropy Principle (MAXENT) is used [11, 12]. The MAXENT is used in multiple fields, including graphics computer, geometric modelling, image processing and supervised learning. This shape function were obtained for Sukumar that established the relation between the convex problem of the maximum-entropy (maxent) statistical inference and the meshfree basis functions with first order consistency [13, 14]. In the context of the Galerkin-based meshfree methods these basis functions were used for compressible and near-incompressible elasticity problems [15]. The main interesting features of the maxent shape functions for SF, is the fact that their values are always positive and smooth, in addition to have a reducing property on the boundary of the domain[16]. Others secondary features, described in [17] and not less relevant, are the variation diminishing property (this mean that the approximation do not create an extreme value) and the monotonicity property. These things are important because at least the first positivity criterion exposed in [18] for shape functions used on SF is fulfilled, situation that not happen in the case of traditional Weight Least Square (WLS). To prove the advantage of this alternative, in this work we used the maxent shape function in a SF meshless method, using the formulation for first and second order consistency proposed in the works of [14] and [17] respectively. We illustrate the performance of the methodology by mean two examples. The first is an one-dimensional PDE value problem with Dirichlet and Neumann boundary condition, and the second corresponds to a diffusion PDE with multiples sources in a 2D-domain.

The remainder of this paper is structured as follows: in section 2, a introduction to maximum-entropy shape functions is presented together with their use in a collocation scheme. Numerical examples are presented in section 3 for two benchmark problems that showing the improved in the numerical approximation in the context of collocation scheme. Finally, in section 4 some final remarks are given.

## 2 MAXENT SHAPE FUNCTION USED IN A COLLOCATION SCHEME

An approximation to an unknown function \( u(x) \) defined in a closed domain \( \Omega \subset \mathbb{R}^d \), which is discretized by a set of field nodes \( x_I \in I = (1, ... N) \), with \( N \) the total field nodes, can be developed considering a linear combination in the form,

\[
u(x) = \sum_{I=1}^{N} \phi_I(x) u_I \tag{1}\]

where \( \phi_I(x) \) and \( u_I \) are known as a shape function and nodal value respectively, associated with the field node \( x_I \). In order to obtain a “local” approximation for function \( u(x) \) the domain \( \Omega \) can be divided into subdomains \( \Omega_I \) that cover the total domain.
domain and to do the approximation in these local regions. Each field node has its local domain, and with the approximation in these local domains is possible to form an equation system whose solution corresponds to the approximated value of the function analyzed in the field node. In a collocation scheme the value are calculated only for the location of the node. For a better understanding of collocation scheme, is possible to review [3]. In the context of the collocation scheme the standard for the shape function are the “WLS” and the “Radial Basis Function” (RBF), but in the present work, these are changed for a new alternative that comes from the statistical mechanic. 

In [11, 12] Jaynes postulated the principle of maximum entropy using for that the entropy of Shannon [19], as a rationale means least-biased statistical inference when insufficient information is available. This principle allows to find the least-biased probability distribution when there are fewer constraints than unknowns in the problem. This led to established that the best solution for the this problem can be obtained by the solution of a convex optimization problem considering the restrictions of the problem environment. This problem was common to many disciplines and then their used is now not only for the information theory, and have applications in various topic as graphical techniques, numerical estimation and prediction, design and analysis of computer experiments, and now used in the Finite Element Method and Meshfree Methods. In this case the connection between the meshfree methods and the maximum entropy (maxent) principles was established by Sukumar in [13]. For that, Sukumar considered that the basis functions \( \phi_I, I \in [1, ..., N] \) associated with the field nodes in the discretized domain can corresponds to the probability distribution \( p_I, I \in [1, ..., N] \) described in the maxent principle [20].

The connection that Sukumar established was considering that the partition of unity and a linearly completeness (condition of a barycentric coordinates), only can prescribe a unique shape function if the polygon region has three nodes. For any other number of nodes region there are multiple possibilities for the shape function. In this context is clear that the maxent can be a good methodology to select the least-biased shape function considering only the real information present in the problem. This was presented for Sukumar in [13], where is obtained a linearly complete interpolants on convex polygonal domains, deriving these from a constrained optimization problem. Next to this in [16] the maxent principle was used to construct convex local meshfree approximation schemes. Here Arroyo demonstrate that the convex optimization problem for a compact set of nodes only will have a unique solution if the region where the nodes are contained is convex. The methodology generalization was realized in [14] taking for that the global and local approach. In [17] and [21] the second order consistency were presented in two different form of the convex problem.

2.1 First Order Formulation

The first order maxent shape function will be presented in the generalized form established in [14]. In this formulation is necessary to define the prior weight function...
denoted by \( m_i(x) \), that corresponds to initial estimate and gives the local character to the shape function. Then to obtain the set of maxent shape function \( \phi_i, i \in \{1, ..., N\} \) is necessary to solve the optimization problem considering the relative maximum entropy functional \( H \) as,

\[
\max H(\phi_1, \phi_2, ..., \phi_n) = -\sum_{i=1}^{N} \phi_i(x) \ln \frac{\phi_i(x)}{m_i(x)}, \text{ subject to }
\begin{align*}
\sum_{i=1}^{N} \phi_i(x) &= 1 \\
\sum_{i=1}^{N} \phi_i(x) \bar{x}_i &= 0
\end{align*}
\] (2a)

An example of prior weight function is the Gaussian radial basis functions represented in equation 3.

\[
m_i(x) = \exp(-\beta_i \|x_i - x\|^2), \quad \beta_i = \frac{\gamma}{h_i^2}
\] (3)

where \( \gamma \) is the parameter that controls the support size of the Gaussian basis functions at node \( i \), and \( h_i \) corresponds at the characteristic length relative to node \( i \). If this prior is used, the convex problem obtained is the same that is established in [16]. The control of the local shape function is then subject to the selection of \( m_i(x) \). When this function has a value of 1, the total domain is used.

In the equation 2b, \( \bar{x}_i = x_i - x \) are the shifted nodal coordinate. It is possible to find the solution of this problem if it is transformed in its dual equivalent problem such as is shown in [16]. This solution will be unique if the domain considered is convex. More information about this procedure can be reviewed in [22]. The first step to obtain the shape functions is to find the Lagrange multiplier vector \( \lambda \) by minimizing \( \ln Z \), where \( Z \) is known as the partition function in statistic mechanic. The nonlinear equations system obtained then, is

\[
\nabla \ln Z = -\sum_{i=1}^{N} \phi_i \bar{x}_i = 0
\] (4)

where \( \nabla \) in this case corresponds to the gradient with respect to \( \lambda \). This nonlinear system can be solved by any numerical method, for example the Newton-Raphson, steepest descent, or any interior-point method. Once \( \lambda \) is completely determined, is possible to obtain the shape functions by,

\[
\phi_i(x) = \frac{Z_i(x, \lambda)}{Z(x, \lambda)}
\] (5a)
\[ Z(x, \lambda) = \sum_{k=1}^{N} Z_k, \quad Z_i(x, \lambda) = m_i(x) \exp(-\tilde{x}_i \lambda^T(x)) \]  

(5b)

where \( \tilde{x}_i = [\tilde{x}_1 \ \tilde{x}_2 \ \tilde{x}_3] \) and \( \lambda = [\lambda_1 \ \lambda_2 \ \lambda_3] \). The gradient of this basis functions is [23],

\[
\nabla \phi_i(x) = \phi_i \left( x_i \left( H^{-1} - H^{-1} \sum_{j=1}^{N} \phi_j \tilde{x}_j \otimes \frac{\nabla m_j}{m_j} \right) + \frac{\nabla m_i}{m_i} - \sum_{j=1}^{N} \phi_j \frac{\nabla m_j}{m_j} \right)
\]

(6)

with \( H \) the Hessian Matrix,

\[
H = \sum_{j=1}^{N} \phi_j \tilde{x}_j \otimes \tilde{x}_j
\]

(7)

### 2.2 Second Order Formulation

The second order maxent basis functions were developed in [17] and [21], and here is only shown the second order shape function. Each of the second order convex shape function \( \psi_i(x) \) will be determined by the convex problem,

\[
\max H(\psi_1, \psi_2, ..., \psi_n) = -\sum_{i=1}^{N} \psi_i(x) \ln \psi_i(x), \text{ subject to } \left\{ \begin{array}{l}
\sum_{i=1}^{N} \psi_i(x) = 1 \\
\sum_{i=1}^{N} \psi_i(x) \tilde{x}_i = 0 \\
\sum_{i=1}^{N} \psi_i(x) \tilde{x}_i \otimes \tilde{x}_i = G(x)
\end{array} \right. \]

(8b)

and the unique solution for this problem is given by,

\[
\psi_i(x) = \frac{\exp(\lambda(x) \cdot \tilde{x}_i + \mu : (\tilde{x}_i \otimes \tilde{x}_i - G(x)))}{Z(x, \lambda(x), \mu(x))}, \text{ where } i \in (1, \ldots, N)
\]

(9)

and

\[
[\lambda(x), \mu(x)] = \arg \min_{\lambda, \mu} \ln Z(x, \lambda, \mu)
\]

(10)
3 NUMERICAL EXAMPLES

In this section are presented an one-dimensional case and other two-dimensional problem that will be analyzed by the use of a collocation scheme using the maxent shape function. To compare this shape function in a strong formulation method, the solution is shown together a traditional Weight Least Square (WLS). The numerical error is estimated by means of a global error defined by

\[
L_2 = \sqrt{\frac{\sum_{i=1}^{N} (u_{i,\text{Exact}} - u_{i,\text{Numerical}})^2}{\sum_{i=1}^{N} (u_{i,\text{Exact}})^2}}
\]

\[
H_2 = \sqrt{\frac{\sum_{i=1}^{N} (u_{x,i,\text{Exact}} - u_{x,i,\text{Numerical}})^2}{\sum_{i=1}^{N} (u_{x,i,\text{Exact}})^2}}
\]

where \( N \) denotes the total number of points in the analysis domain and its boundary, and \( u_{x,i,\text{Numerical}}, u_{x,i,\text{Exact}} \) represent the solutions at each point obtained numerically and exact, respectively.

3.1 One-dimensional Poisson equation

A one-dimensional Poisson equation with high gradient in a local area is analyzed as the first example. This one-dimensional problem is expressed mathematically by the following governing equation and its boundary conditions,

\[
\frac{d^2 u}{dx^2} = -6x - \frac{2}{\alpha^2} \left( 1 - 2 \left( \frac{x - \beta}{\alpha} \right)^2 \right) \exp \left[ -\left( \frac{x - \beta}{\alpha} \right)^2 \right], \quad 0 < x < 1 \quad (12a)
\]

\[
u(x = 0) = \exp \left[ -\left( \frac{\beta}{\alpha} \right)^2 \right], \quad (12b)
\]

\[
\left. \frac{du}{dx} \right|_{x=1} = -3 - \frac{2}{\alpha^2} (1 - \beta) \exp \left[ -\left( \frac{1-\beta}{\alpha} \right)^2 \right] \quad (12c)
\]

The exact solution of the above Poisson Equation is given by,

\[
u(x) = -x^3 + \exp \left[ -\left( \frac{x - \beta}{\alpha} \right)^2 \right] \quad (13)\]
The parameters of $\alpha$ and $\beta$ used in the present analysis are 0.05 and 0.5 respectively.

In the case of uniform points presented in the figure 1 and 2, corresponding to the function and their derivative respectively, it is possible to appreciate the best performance of the maxent shape function over the traditional WLS. This situation is appreciable to the first and second order maxent in the figure 1b and 2b. The convergence of the proposed alternative is clear.

In the case of non-uniform points presented in the figure 3 and 4, corresponding to the function and their derivative respectively, again it is possible to appreciate the best performance of the first order maxent shape function over the traditional WLS.
3.2 Poisson problem with localized sources

This example shows the convergence in the $L_2$ estimator of the collocation scheme using first order maxent shape functions to approximate a PDE. In particular, this example shows how the proposed formulation can deal with localized sources. For this purpose is considered the Poisson boundary value problem in the form presented in [21],

\begin{align}
-\nabla \cdot (k \nabla u) &= s(x, y), \quad \in [0, 1] \times [0, 1], \\
u &= \bar{u} \quad \in x = 0, x = 1, y = 0, \\
\nabla u \cdot n &= 0 \quad \in y = 1.
\end{align}

where $s(x, y)$ and $\bar{u}$ correspond to the second derivatives and the solution $u(x, y)$ evaluated on the boundary respectively. In this case $k = 1$. 
Table 1: Coefficients to calculate the exact solution of the problem.

<table>
<thead>
<tr>
<th>i</th>
<th>A_i</th>
<th>β_i</th>
<th>x_i</th>
<th>y_i</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>180</td>
<td>0.51</td>
<td>0.52</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>450</td>
<td>0.31</td>
<td>0.34</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>800</td>
<td>0.73</td>
<td>0.71</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td>1000</td>
<td>0.28</td>
<td>0.72</td>
</tr>
</tbody>
</table>

\[ u(x, y) = \sum_{i=1}^{4} A_i \exp(-\beta_i((x - x_i)^2 + (y - y_i)^2)) \] 

Similar to the case of one-dimensional Poisson equation, the numerical results show that the approximation tends to converge to the exact solution in a low number of uniform nodes. We try to compare with the traditional methods (WLS), but the accuracy of that failed in a great order with respect to the proposed alternative and for this reason we only presented in table 2 a comparative value of the error for the same discretization. Beside to this, in figures 5 and 6 the behavior of the error estimator and the approximation solution using maxent alternative is shown. In both figures are possible to establish that the method tends to converge to the exact solution of the problem with rate that is considered good for a collocation scheme.

![Figure 5](image_url)

(a) Logscale for L2 and h.  
(b) L2 vs Number of points.

**Figure 5:** Estimator behavior.
Table 2: Comparison between estimator error for Maxent using $\gamma$ 4.0 and WLS using $\alpha$ 3.0

<table>
<thead>
<tr>
<th>N Puntos</th>
<th>$L_2$ - Maxent</th>
<th>$L_2$ - WLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>49</td>
<td>1.87994292E+01</td>
<td>8.43571762E+02</td>
</tr>
<tr>
<td>121</td>
<td>3.04702533E+00</td>
<td>2.88033198E+02</td>
</tr>
<tr>
<td>441</td>
<td>2.55132938E+00</td>
<td>2.44358092E+03</td>
</tr>
<tr>
<td>625</td>
<td>3.54516989E-01</td>
<td>9.56362982E+01</td>
</tr>
</tbody>
</table>

(a) 40x40 Approximation.  (b) 40x40 Theoric.

Figure 6: Comparison between approximation and theoric solution.

4 CONCLUSIONS

In this article, a meshfree collocation method using maxent shape functions was presented. This shape functions have interesting features as their positivity and smooth, that are requirements desired when a collocation scheme is used. With this is possible to have an alternative to traditional WLS shape function. Numerical examples of one-dimensional and two-dimensional problems was presented to demonstrate the efficiency of the method. Based on this study, is possible established the following conclusion,

- The numerical results obtained by this method show best behavior that the tradional WLS in regular grid and a equal or best behavior in irregular grid. However, further studies are needed to improve the performance of the method, focusing efforts in the derivatives of maxent shape function on the boundary, because here is not always simple to calculate the value. The method shows an stability and robustness that is not present in the traditional option. Moreover, further studies on the calculation accuracy and convergence rate, including three-dimensional case, are needed.

- This type of shape function not needed any connection between nodes in the whole
calculation process, and hence is a truly meshfree method. This shape function, too,
is a result of a convex optimization problem, and this implies that the approximation
is the best for the real information present in the problem.

- The computational cost is similar to the traditional WLS inclusive for not uniform
distribution.

5 ACKNOWLEDGEMENTS

The authors acknowledge the support of this work by the Aula UTFSM-CIMNE, by
the DIMEC-UTFSM and by the DGIP-UTFSM.

References


and integration schemes for meshfree methods”. International Journal for Numerical

integration for Galerkin meshfree methods”. International Journal for Numerical


630, 1957.


A TRADE-OFF ANALYSIS BETWEEN HIGH-ORDER SEISMIC RTM AND COMPUTATIONAL PERFORMANCE TUNING

DANILO L. COSTA¹, ALVARO L. G. A. COUTINHO¹, BRUNO S. SILVA², JOSIAS J. SILVA² AND LEONARDO BORGES³

¹ High Performance Computing Center, COPPE/UFRJ
PO Box 68506, Rio de Janeiro, RJ 21945-970, Brazil
alvaro@nacad.ufrj.br
www.nacad.ufrj.br

² Petrec, Petroleum Research and Technology
R. Hélio de Almeida, s/n - Cidade Universitária, Rio de Janeiro, RJ, Brazil
josias@petrec.com.br
www.petrec.com.br

³ Intel Corporation
Software and Services Group, USA
leonardo.borges@intel.com
www.intel.com

Key words: Reverse Time Migration, High-Order FD Operators, Computational Optimizations, Intel Xeon Phi.

Abstract. This work aims to examine the consequences of the variation of the FD operator order over the computational cost of the 3D Acoustic Isotropic RTM kernel. Both the Taylor operators and the operators optimized by the Binomial Window are assessed, whose discretization orders range from 4th to 36th. The analysis presented in this paper were performed on two computing platforms - Intel Xeon Phi 7120P coprocessor and Intel Xeon E5-2697v2 processor. Since the RTM’s performance is strongly influenced by architecture aspects such as cache size, memory bandwidth and processor’s clock (Krueger et al., 2013), performance profiling was made over each platform in order to evaluate hardware efficiency and perform computational optimizations. Therefore, this paper aims to establish possible relations between the order of the operators and the performance of each architecture, additionally intend to assists in the decision-making process upon choosing the finite difference discretization order. Considering the used dataset and seismic parameters, experiments have shown that the optimized 16th order operator offers the best performance. This one was able to save more than 70% in processing time and saves up to 90% in memory demand. The results also showed the choice of the order of the FD operator should consider aspects such the number of flops (floating-point operations) at the stencil computing, hardware architecture, domain size and shape.
1 INTRODUCTION

Among the several methods used in the Oil & Gas geophysical exploration process, the Seismic Method is considered to be the most important one. Such method is based in the propagation, in the sub-surface, of seismic waves produced by artificial means, and its main purpose is to outline the geological structures of the Earth’s sub-surface. The problem with the seismic wave propagation simulation consists of solving differential equations that describe the propagation inside the Earth, under a set of initial, final and boundary conditions. Among the several numerical methods proposed to solve such partial differential equations, the Finite Difference (FD) method is one of the most successful. And the Reverse Time Migration (RTM) is one of the main approaches in the seismic processing industry for imaging the subsurface structure of the Earth.

However, Finite Element and Spectral Methods - or combinations of both - also are often employed in seismology. Whatever the method used, it is essential to efficiently implement applications that explores architectural features of modern processors. Nowadays, high performance computers achieve processing Petaflops (10^15 floating-point operations per second, or flops) capacities. Usually these machines are composed of heterogeneous compute nodes constituted by multi-core processors connected to many-core coprocessors or to graphic accelerators. In turn, the nodes are interconnected by high-speed infiniband networks.

Some estimates predict that machines able to reach Exaflops (10^18 flops) will be available within 5 years. To this end a huge effort to reduce the energy demand of these machines, attention to fault tolerance and investment in new programming models will be necessary. Considering that the seismic processing should be ready for this new computing paradigm, this paper aims to point out new trends in processors architecture that improve computational efficiency and which may significantly impact seismic processing in a near future. On the other hand, the increase in performance does not result solely from the use of powerful hardware, but also comes from algorithmic innovations. Thus, this paper also discusses some possible algorithmic innovations impacts over RTM performance. The remainder of this paper is organized as follow. Next section presents a brief RTM computational cost analysis. Section 3 introduces improved finite differences schemes, while Section 4 shows the numerical experiments. The paper ends with a summary of our main conclusions.

2 RTM COMPUTATIONAL COST ANALYSIS

The RTM computational costs usually are quite high either in relation to runtime or in relation to hardware resources demand. In turn, seismic modeling using the full wave-equation constitutes the RTM's kernel which is the key source of most of computational costs and thus should be the optimization main target. Consequently, all optimizations and discussions performed in this work concern the 3D acoustic isotropic full wave-equation stencil. However, the same strategies were used to optimize more complicated algorithms, such as Transverse Tilted Isotropy (TTI) RTM.

A usual strategy to decrease the computational cost consists in discretize the wave-equation in space and time by a FD scheme, in which high-order Taylor coefficients are used to approximate the spatial differentials operators. This strategy could enlarge distances between mesh points and thus decreases both the memory amount needed to store the mesh and the
spatial-iterations needed to update the wave field. In order to maximize these benefits, the conventional Taylor coefficients can be replaced by optimized FD coefficients.

Optimized coefficients are employed to improve numerical accuracy or, since enough precision has been achieved, may be used to enforce even larger grid spaces. Unfortunately, either using optimized coefficients or high-order Taylor coefficients there is an increase in the amount of time-iterations needed to propagate the wave field throughout the whole spatial domain. There is also an increase in the amount of arithmetic operations in the wave-equation stencil enhancing the computational intensity. These latter statements have potential to reduce the benefits previously achieved and, therefore, justifies an analysis that aims a trade-off between computational performance and the choice of appropriate FD coefficients.

3 IMPROVED FD SCHEME

Liu and Sen\(^6\) have proposed a FD scheme whose derivatives coefficients fully satisfies the numerical dispersion analysis for a given wave field frequency. Then Liu and Sen\(^7\) applied the time-space operator to the acoustic wave-equation. This approach gives a stencil with more precision than a stencil obtained by a conventional Taylor expansion. In order to achieve further precision, Liu\(^8\) has developed a scheme to get optimized FD coefficients using the least squares, trying to minimize the numerical dispersion error for both the second-order spatial derivative and the time-space derivatives.

The conventional FD operator for a second-order spatial derivative in a given function \(f(x)\) can be approximated by a Taylor expansion around the point \(x=0\) (Equation 1).

\[
\frac{\partial^2 f}{\partial x^2} \approx \frac{1}{\Delta x^2} \sum_{n=-N/2}^{N/2} a_n \left[-\frac{2}{n^2} \cos(n \pi)\right] f(\Delta x),
\]

where \(\Delta x\) is the sampling interval along the \(x\) axis, \(N\) is the discretization order and \(a_n\) coefficients are defined by the binomial window in Equation 2.

\[
a_n = \binom{N}{N/2+n} / \binom{N}{N/2}
\]

A way to reduce the numerical dispersion is to adopt window functions that generates optimized coefficients for Equation 1. Chu and Stoffa\(^9\) have proposed a window \(w_n\) in order to obtain optimized FD coefficients which may be expressed as shown in Equation 3.

\[
\frac{\partial^2 f}{\partial x^2} \approx \frac{1}{\Delta x^2} \sum_{n=-N/2}^{N/2} w_n \left[-\frac{2}{n^2} \cos(n \pi)\right] f(\Delta x),
\]

\[
w_n = \frac{n + N + M}{N^2/2 + n}
\]
The optimized window function \( w_n \) that appears in Equation 3 can be written as in Equation 4. The final form of the optimized FD operator (Equation 5), can be held replacing the previous window function \( w_n \) in Equation 3, that is,

\[
\frac{\partial^2 f}{\partial x^2} \approx \frac{1}{\Delta x^2} \sum_{n=-N/2}^{N/2} c_n f(\Delta x),
\]

Derivatives coefficients \( c_n \) final form may be then defined either for the conventional Taylor operator (Equation 6) or for the optimized binomial window operator (Equation 7 and Equation 8), that is,

\[
c_n = a_n \left[ -\frac{2}{n^2} \cos(n \pi) \right]
\]

\[
c_n = w_n \left[ -\frac{2}{n^2} \cos(n \pi) \right]
\]

\[
c_0 = \sum_{n=-N/2}^{N/2} c_{-n} + c_n
\]

Establishing the \( M \) parameter in Equation 3 is a non-trivial issue to address, to be handled carefully since it can leads to disastrous changes in the calculation results. Zhang and Yao have determined coefficients \( c_n \) by using an optimization scheme which maximizes the wave number convergence within certain error limitations. In order to establish the optimized coefficients in this paper the Simulated Annealing (SA) minimization algorithm was employed. The absolute spectral error expressed in Equation 9, that normalizes the wave number \( k_x \) by the Nyquist wave number \( k_{xn} \), was used to perform a spectral error analysis for several discretization orders.

\[
Error \left( \frac{k_x}{k_{xn}} \right) = \left[ c_0 + 2 \sum_{n=1}^{N/2} c_n \cos \left( \frac{k_x}{k_{xn}} n \pi \right) \right] - \left( \frac{k_x}{k_{xn}} \pi \right)
\]

Figure 1: spectral error analysis performed to second-order spatial derivative.
Analysis results for the conventional operator and for the optimized operator can be found together in Figure 1. As expected, results reinforce that the optimized binomial window operators are able to achieve more precision than conventional Taylor operators. For instance, the optimized 16th order operator is much more accurate than the conventional 32th order operator.

4 NUMERICAL EXPERIMENTS

From a practical viewpoint, it is important to assess the advantages and disadvantages of the high-order FD approaches taking into account computational architecture aspects. Such relationship may indeed be useful to clarify seismic modeling behavior under three main aspects: (i) memory demands; (ii) processor performance; (iii) and seismic modeling runtime.

Despite algorithm implementation features, the first aspect has a strong dependency on the discretization order either using optimized coefficients or Taylor coefficients, as shown in Figure 2. For both of them there is a substantial reduction in memory demands with the increase of discretization order, though for optimized operators this reduction is more pronounced. Note that the optimized 16th order can be more efficient than the Taylor 36th order, and saves up to 90% in memory usage compared to Taylor 4th order.

The analyses presented in this paper were performed on two computing platforms: Intel Xeon Phi 7120P coprocessor and double Intel Xeon E5-2697v2 processor. As stated before, seismic modeling performance change in function of certain features of these platforms such as processor frequency, number of cores, memory bandwidth and so on. Thus, performance profiling tools were used to investigate these relationships and then high performance computing (HPC) techniques were applied in order to improve algorithm performance on such architectures.

Based on profiling analysis on both platforms, the main HPC techniques performed in RTM algorithm are parallelization, vectorization, thread affinity, memory alignment, padding, prefetching, loop unrolling (just for Xeon E5) and cache blocking (just for Xeon Phi). Figure 3 shows, in sequence, the performance improvement provided by each optimization. Should be highlighted the importance of ensuring the effective association between parallelization (Vs01)
and vectorization (Vs02). Despite the high scalability of the loop parallelization, ensuring the effectiveness of this association is the first big optimization effort and is crucial to extract high level of performance on current massively parallel processors.

Since the algorithm is optimized for the two platforms we are able to proceed with the tests involving the several wave-equation discretization orders for both FD coefficients approaches. Performance behavior shown in Figure 4 come not only from the architectures features and optimizations performed but also come from the stencil arithmetic intensity and spatial mesh size for each discretization order. Combining these last two factors is a critical point for data traffic into memory hierarchy.

Note that performance does not grow indefinitely with the increase of discretization order, so there is an optimal order for each platform. For the Intel Xeon E5-2697v2 processor greater efficiency occurs for Taylor 14th order. For the Intel Xeon Phi 7120P coprocessor maximum performance takes place for the optimized 16th order, combining the hardware peak efficiency.
with minimal memory demand. This fact is quite convenient considering the limited amount of embedded memory in Xeon Phi coprocessor.

We still need, however, to verify the effects of the coefficients and approximation order on the runtime, after all, the overall processing time encompasses all aspects of the computational cost. Figure 5 shows the overall processing time for both FD coefficients approaches.

![Figure 5: Runtimes for 0.5 sec of seismic acquisition for different architectures, coefficients and FD order.](image)

For the Intel Xeon Phi 7120P coprocessor minimum runtime takes place for the optimized 16th order and agrees with the results shown in Figure 3 and Figure 4. On the other hand, the lowest processing time for Xeon E5 processor does not happen for the Taylor 14th order, instead, the lowest value is achieved by the optimized 14th order. Ultimately, due the similarity between their values and aiming to save more memory, is appropriate to assume that optimized 14th and 16th orders have equivalent runtime. Indeed, the advantages provided by the great reduction in memory demand overcome the disadvantages coming from rises in both the number of time-iterations and in the stencil arithmetic complexity.

5 CONCLUSIONS

Regarding the choice of an optimal discretization order, the present results show that it must be considered factors such as computational complexity, processor architecture, optimizations performed and modeled domain dimensions. Under the circumstances presented here, it is feasible to state that the optimized Binomial Window 16th order operator provides higher efficiency on the two tested architectures. Allowing to cut down the seismic modeling runtime in 77% and saving up to 90% in memory usage compared to Taylor 4th order.

Such savings can be even more relevant considering the limited amount of embedded memory on modern massively parallel architectures, such as the Intel Xeon Phi coprocessor and graphical accelerators (GPUs). This control of the required amount of memory can also be applied in future generations of many-core architectures that will support multilevel main memory with different sizes and access speeds - such as the flat memory model on the Xeon Phi future generation. In addition, the pressure decrease over computing resources promoted by the high-order optimized operators is also very useful when applied to the TTI RTM 3D migration, whose computational demands are still higher than those in acoustic isotropic RTM.
REFERENCES


A NEW PARALLEL NUMERICAL MODEL FOR MULTIPLE COLLISIONS

ROBERTO LEON$^{1,3}$, LUIS SALINAS$^{1,2}$, CLAUDIO TORRES$^{1,2}$ AND GONZALO HERNANDEZ$^2$

$^1$Departamento de Informática, Universidad Técnica Federico Santa María
Avenida España 1680, Valparaíso, Chile

$^2$CCTVal, Universidad Técnica Federico Santa María
Avenida España 1680, Valparaíso, Chile

$^3$Facultad de Ingeniería, Universidad Nacional Andrés Bello
Quillota 980, Viña del Mar, Chile

Key words: computational dynamics, multiple collisions, pseudopotential of dissipation, rigid body rotation

Abstract. The A-CD$^2$ method [8] is used for numerical simulation in collisions of rigid bodies. This method solves a large scale constrained minimization problem in order to obtain the new velocities of each particle after a collision. The minimization problem is originally solved using the Uzawa method [6]. Unfortunately, the Uzawa method converges very slowly. In order to overcome this shortcoming, we propose a new numerical model that improves the previous approach in three ways: (i) the minimization problem is solved with an interior point method [10, 23] implemented in the CPLEX library [7], which converges in half of time required by the Uzawa method; (ii) the new model includes the rotational dynamics of the free moving rigid bodies between collisions; (iii) a parallel implementation of the updated velocities after collision is included. This new approach allow us to obtain numerical simulations with a large number of solids that was not possible before.

1 INTRODUCTION

Collision between rigid bodies is an important class of process arising in several natural events like avalanches, molecules and asteroids collisions, etc. It also arises in many important engineering and industrial problems e.g., mining, construction, transport, etc.

Thus, it is important to develop computational models for collision simulations in order to avoid experiments, which usually need a lot of resources and equipments and hence are expensive. A computational simulation allows to considerably reduce the use of special
equipments and other resources, and still allowing to obtain a very close description of the dynamics of the phenomenon, with the goal of analyzing and studying the evolution of collisions in time.

In this work, following Fremond’s approach [14], we address the collision problem as a system deformation in the sense that a set of rigid bodies constitutes a system, which is deformable due to changes in the distance between the bodies [14]. Much research is based on this theory, e.g., collisions of balls [5], crowd movements [21], collisions and fractures [2, 3, 13, 16], collisions in fluids [17], collision in granular flow [8, 9, 11, 12], among others.

There are of course other approaches –different from Fremond’s method– to deal with similar problems, e.g., impact between rigid bodies [1], fall of rigid bodies [24], particle collision [19], nonlinear deformable bodies [4], flows with collisions [25], among others.

The A-CD$^2$ method –the Atomized stress Contact Dynamics respecting a Clausius-Duhem inequality method– is presented in [8]. This approach assumes constant linear and angular velocity of the solids in the time interval between two consecutive time-steps. This is clearly a non physical description of the movement of bodies in those time intervals. In order to overcome this shortcoming we introduce in this paper a new model, which takes into account the physical dynamics of the movement of the bodies in the time intervals between two consecutive time-steps. In particular, neither the linear speed nor the angular speed of the solids are assumed to be constant during those time intervals.

From the computational point of view, the A-CD$^2$ approach uses the Uzawa method to solve a constrained minimization problem. This method is highly non-efficient due to the huge amount of iterations (over 1000 per timestep) necessary to obtain a solution. In this paper we replace Uzawa’s method by a the interior point method included in the CPLEX library [7]. With these changes the computational time decreases 3 orders of magnitude, opening the possibility to simulate a larger amount of solids ($\geq 2000$).

2 A-CD$^2$ METHOD

The Atomized stress Contact Dynamics respecting a Clausius-Duhem inequality method (A-CD$^2$) [8], is summarized by the following three recursive steps: (1) solids move with constant velocity during $\Delta t$, (2) contact points between solids are computed at the end of each time step $\Delta t$, and (3) at the end of each simulation step, the new velocities for the next $\Delta t$ are computed according to the dynamics of the experiment.

In order to obtain the velocities after the collision, we use instantaneous collision model proposed by Dimness [11]. For the sake of simplicity [11] we first consider a single point of mass $m$, colliding once with a fixed rigid horizontal plane at some instant $t_c$ in the time interval $[t_k, t_{k+1}]$, ($\Delta t = t_{k+1} - t_k$) and having instantaneous velocity discontinuity at $t_c$. Then we generalize the formulation to simultaneous collision of $N$ solid rigid rectangular boxes [11].

Collisions are assumed to be instantaneous, being the velocity $\vec{u}$ of the point discontinuous at the instant $t_c$ of the impact. The left and right limits of the velocities are denoted
by $\vec{u}^-$ and $\vec{u}^+$, respectively. The contact force is concentrated in time, becoming a percussion denoted by $\vec{P}^{int}$ [14]. In figure 1 we see a interaction between a point and a fixed rigid plane, specifically we see the description of the point trajectory $\vec{x}(t)$ in the interval $[t_k, t_{k+1}]$ and the collision at time $t_c$.

![Figure 1: Point - plane system: a particle with mass $m$ and a fixed plane.](image)

Interior forces are considered in the particle motion at interval time $[t_k, t_{k+1}]$ and contact forces $\vec{f}^{int}$ are considered as a internal percussion $\vec{P}^{int}$ concentrated in time. Since the particle velocity is not continuous at $t_c$ [11], we consider a virtual velocity of deformation at the discontinuity.

In the case of a contact, constitutive laws describe the interactions among particles during the collision and have to assure the non-interpenetration of the solids. Internal percussion is therefore divided in two parts, a dissipative percussion $\vec{P}^d$ and a reactive percussion $\vec{P}^{reac}$ [8, 11]

$$\vec{P}^{int} = \vec{P}^d + \vec{P}^{reac}$$  \hspace{1cm} (1)

Dissipative percussion describes the (dissipative) interactions among colliding solids. In general, dissipative interaction associated can be described introducing, a convex, positive and null in the origin pseudopotential of dissipation $\Phi^d$ [11, 15, 20, 22].

Reactive percussion describes the non-interpenetration condition, which implies $u_N^+ = \vec{u}^+ \cdot \vec{N} \geq 0$ (Figure 1). This percussion is null if $u_N^+ > 0$, is active if $u_N^+ = 0$ and implies that the condition $u_N^+ < 0$ can not be verified. All these properties can be written by means of the indicator function. The internal percussion can be written in the following form

$$\vec{P}^{int} \in \partial\Phi \left( \frac{\vec{u}^- + \vec{u}^+}{2} \right) \quad \text{where} \quad \Phi = \Phi^d + 1_K$$

$$K = \left[ \frac{\vec{u}^- \cdot \vec{N}}{2}, \infty \right]$$  \hspace{1cm} (2)
According to the assumptions made above, the problem to solve at an instant of the velocity jump is considered as a simultaneous collision of several solids. The following minimization problem obtains the velocities after collision \[8, 11\]

\[
\inf_{\vec{x}} \langle \vec{x}, \vec{x} \rangle + \Phi(\vec{x}) - \langle 2\vec{U}^- + \vec{\tau}^{ext}, \vec{x} \rangle
\]  

(3)

The collision problem is now reduced to a constrained minimization problem. The minimization problem considers the dissipative percussion \(\Phi_d\) as a quadratic function and the non-interpenetration condition as a set of constraints \(\Omega = \varphi_k(x) \leq 0, \ k = 1, n_c\) which determines the reactive percussion of the internal percussion.

The constrained minimization problem is given by:

\[
\inf_{x \in \Omega} F(x) = \langle x, x \rangle + \Phi_d(x) - \langle 2\vec{U}^- + \vec{\tau}^{ext}, x \rangle
\]

where \(\Omega = \{ x \in \mathbb{R}^{6N} | \varphi_k(x) \leq 0, \ k = 1, n_c \}\)  

(4)

3 FREE RIGID BODY MOTION

The A-CD\(^2\) method considers a constant velocity between two timesteps. This could be improved by considering the equations from classical mechanics [18]:

\[
\sum \vec{F} = m \vec{a}
\]

(5)

\[
\sum \vec{M_G} = \vec{H_G}
\]

(6)

where \(\sum \vec{F}\) is the sum of all external forces acting on the solid, \(\sum \vec{M_G}\) is the sum of all external momenta applied to the solid and \(\vec{H_G}\) is the angular momentum of the solid about its mass center. From (6) is possible to obtain the Euler equations of motion concerning the rotation motion:

\[
\sum M_1 = I_{11} \dot{\omega}_1 + (I_{33} - I_{22}) \omega_2 \omega_3
\]

\[
\sum M_2 = I_{22} \dot{\omega}_2 + (I_{11} - I_{33}) \omega_1 \omega_3
\]

\[
\sum M_3 = I_{33} \dot{\omega}_3 + (I_{22} - I_{11}) \omega_1 \omega_2
\]

(7)

where \(I_{11}, I_{22}\) and \(I_{33}\) are the moments of inertia in the direction of the principal axis. In our case, since the motion is torque-free, the Euler equations gives:

\[
\vec{H_G} = \sum \vec{M_G} = \vec{0}
\]

(8)
which means that the direction (and the magnitude) of $\overrightarrow{H_G}$ remains fixed. In this case [18] equations (7) reduce to:

\[
\begin{align*}
I_{11}\dot{\omega}_1 &= (I_{22} - I_{33}) \omega_2 \omega_3 \\
I_{22}\dot{\omega}_2 &= (I_{33} - I_{11}) \omega_1 \omega_3 \\
I_{33}\dot{\omega}_3 &= (I_{11} - I_{22}) \omega_1 \omega_2
\end{align*}
\]  

(9)

with an initial condition for $\omega_1(0)$, $\omega_2(0)$ and $\omega_3(0)$ provided by the initial states of the solids.

The equations in (9) were solved using the Euler numerical integration method, computing the following set of equations:

\[
\begin{align*}
\omega_1(t + \Delta t) &= \omega_1(t) + \Delta t \left( \frac{I_{22} - I_{33}}{I_{11}} \right) \omega_2(t) \omega_3(t) \\
\omega_2(t + \Delta t) &= \omega_2(t) + \Delta t \left( \frac{I_{33} - I_{11}}{I_{22}} \right) \omega_1(t) \omega_3(t) \\
\omega_3(t + \Delta t) &= \omega_3(t) + \Delta t \left( \frac{I_{11} - I_{22}}{I_{33}} \right) \omega_1(t) \omega_2(t)
\end{align*}
\]  

(10)

Figure 2: Motion of solid rigid rectangular box animated with angular speed and a positive horizontal speed under the action of a vertical gravitational field. The red curve represents the trajectory of the center of gravity of the box. The blue curve represents the trajectory of a vertex of the box. (a) Trajectories with constant angular speed. (b) Trajectories taking into account the inertial dynamics of the solid rigid rectangular through the successive solution of the corresponding Euler equations of the movement.
4 COMPUTATIONAL SIMULATION

Numerical simulations have been performed using the Euler method. The physical time simulated has been 1 second (10000 time steps) for all the numerical experiments. We have used $\Delta t = 10^{-4}$ [s] to satisfy stability constraints of the Euler method. Notice that we have verified that for each time step of all the numerical experiments the eigenvalues of the Jacobian matrix times $\Delta t$ were inside the stability region of the Euler method.

The simulation main algorithm is based in three modules: contact detection, optimization and updating velocities; in Figure 3 shows the interaction between the modules. The numerical code has been developed in C, and performed in the cluster LEVQUE as part of the NLHPC project. This cluster is an IBM iDataplex machine with 536 compute cores. Each node (8 cores) is equipped with two quad-core Intel Xeon X5550 processors running at 2.67GHz with 24GB RAM.

In Table 1 is shown the computational time with A-CD$^2$ method and our new approach (sequential and parallel). It is possible to observe that 1000 solids or greater is not possible to simulate with A-CD$^2$ method. Our approach is capable to simulate until 2000 solids, with the challenge to increase this amount. The parallel time is in average a 10% better than sequential, because certain parts of the code were parallelized, opening the possibility of improving it later.
<table>
<thead>
<tr>
<th>Number of solids</th>
<th>A-CD$^2$</th>
<th>New approach (sequential)</th>
<th>New approach (parallel)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>68.23</td>
<td>2.88</td>
<td>2.59</td>
</tr>
<tr>
<td>100</td>
<td>298.45</td>
<td>6.55</td>
<td>5.69</td>
</tr>
<tr>
<td>500</td>
<td>567.32</td>
<td>85.69</td>
<td>72.84</td>
</tr>
<tr>
<td>1000</td>
<td>–</td>
<td>313.28</td>
<td>285.08</td>
</tr>
<tr>
<td>2000</td>
<td>–</td>
<td>1302.09</td>
<td>1171.88</td>
</tr>
</tbody>
</table>

Table 1: Computational time for A-CD$^2$ method and the new approach, with the sequential and parallel program.

5 CONCLUSIONS

In this short paper we presented a new computational model for multiple collision of rigid bodies. The corresponding simulation uses so far the so called A-CD$^2$ method. Our new approach is based on the CPLEX library and improves the A-CD$^2$ method in three ways:

(i) Both the old and the new method need to solve a large constrained minimization problem. The A-CD$^2$ approach uses the Uzawa method to solve this problem. The new approach uses an interior point method implemented in the CPLEX library and is capable to solve the minimization problem in half of the time required by the Uzawa method.

(ii) The old A-CD$^2$ does not consider the inertia components of the rigid bodies for the movements between collisions. The new model includes the rotational dynamics of the free moving rigid bodies between collisions. This difference is significant for a more accurate simulation of real collision phenomena, specially when the geometry of the colliding rigid bodies imply very different principal inertia moments.

(iii) In our new approach we consider both sequential and parallel implementations for updating linear and angular speeds between collisions.

The computation time is greatly improved (90% in average). The amount of solids to simulate with our new approach is larger than what is possible with the A-CD$^2$ method, opening the chance to consider larger sets of colliding bodies. The parallel computing time is in average only a 10% better than the sequential, because only certain parts of the code have been parallelized. This part will be improved in future implementations. Our next challenge will be to further diminish the computing time, thus allowing the simulation of larger numbers of rigid bodies. Another important task to improve is the contact detection module. A profile analysis showed that edge contact detection between bodies is very expensive. Thus a more efficient edge contact detection algorithm is needed.
6 ACKNOWLEDGMENTS

The authors acknowledge support from the CCTVal (Centro Científico Tecnológico de Valparaíso). Special thanks to NLHPC (National Laboratory for High Performance Computing) project.

REFERENCES


PARALLEL DELAUNAY TETRAHEDRIZATION OF A GIVEN DISCRETE POINT SET CONSTRAINED BY AN IMPOSED BOUNDARY MESH

PABLO J. NOVARA\textsuperscript{1,2} AND NESTOR A. CALVO\textsuperscript{1,2}

\textsuperscript{1}Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral (FICH-UNL)
\textsuperscript{2}Centro de Investigación de Métodos Computacionales (CIMEC-CONICET)
\{pnovara,ncalvo\}@unl.edu.ar

Key words: Tetrahedrization, parallelism, Delaunay

Abstract. This article presents a parallel method for generating a valid finite elements all-tetrahedral 3D mesh, starting from an imposed boundary mesh and a fixed set of points, following the Delaunay criteria whenever possible. The proposed meshing algorithm connects the points inside the volume described by the boundary mesh, building and all-tetrahedral mesh without adding, removing, nor moving input points, and fitting that imposed boundary. This is a very common scenery in particle methods such as PFEM2\cite{6}. The approach combines DeWall\cite{3} algorithm with some techniques inspired in advancing front methods. It can be parallelized, for both shared and distributed memory models. This paper focuses on efficiently solving problems related to boundary meshes that make it impossible to enforce the Delaunay condition, and discusses many implementation details that arise when dealing with threads synchronization, numerical errors, and ambiguous Delaunay configurations, in order to provide a detailed and reproducible algorithm. Results from a shared memory implementation are also presented here, discussing the actual data structures and the required tweaks in order to obtain competitive times, showing a very good stability, scalability, and parallel efficiency.

1 Introduction

Parallel computer architectures can be divided in two main groups: shared memory and local memory. The former refers to architectures where multiple processors share a common RAM memory. In this case, memory access can be a bottle-neck if the tasks are memory-intensive, and special care must be taken in synchronization between threads to avoid common problems in parallel programming such as race-conditions. In a local memory architecture, usually the full problem must be divided into smaller subproblems. Every process have its own memory, and should store everything it needs to solve its subproblem, therefore a lot of processing and communication can be required to generate
and distribute those subproblems and to merge partial results generated on each processor. Also, some data can be replicated across processors, and actions on one process that modify the replicated data in others may need to be broadcasted, requiring extra communication and sometimes generating a sensible overhead. From this perspective, different meshing algorithms can be also classified by how coupled this subproblems are. The coupling level determines the amount of inter-process communication required to solve the task and to merge the results. A highly coupled meshing algorithm will be more suitable for shared memory architectures and will allow solving already solvable problems faster than a serial version, but it might not allow solving bigger problems. A decoupled algorithm is usually preferred, but may require big computational effort to generate the different decoupled subproblems and/or to merge the partial solutions (i.e. to modify common interfaces between subdomains to ensure that they are fully compatible).

Parallel mesh generation algorithms can benefit almost every Computer Fluid Dynamics (CFD) simulation methods. Parallelization can improve CFD simulations in two different ways: can be used to solve bigger problems, and can be used to solve the same problems faster. Bigger problems can be understood as solving problems on bigger domains, and also as solving the same problems that serial solvers do, but with higher precision. This is generally true when solving real problems with classic methods such as Finite Elements Method (FEM) where memory consumption can be a strong limitation. A good local memory parallelization strategy can relax this limitation by allowing the generation of bigger or more dense meshes. The other benefit from a parallel mesh generation algorithm is the increased speed in the overall method. Methods such as the Second Generation Particle Finite Elements Methods, aimed at real-time simulation, find a big bottleneck in the meshing stage required in moving-mesh implementations (PFEM-2[6]). An efficient parallel algorithm for mesh generation can cause a big impact in the overall time for these simulation.

However, particle-based methods impose some special constrains to mesh generation algorithms, thus requiring a different approach. Mesh generation usually involves the generation of inner nodes among with its connections (elements) for a given domain. In particle-based methods, the particles usually are the nodes, so the meshing algorithm goal is to find the best connections for the imposed set of nodes in order to obtain a mesh. For the rest of this article, we will focus on this problem. It is important to note that having the number and position of mesh-nodes given beforehand can be deceptively seen as a simplification (as part of the meshing problem already solved). But it is actually an extra constrain that will limit the set of operations the algorithm can perform to achieve its goal, making it a harder to solve problem.

2 Method outline

As stated before, the problem is to generate a tetrahedrization for a given set of points constrained by an imposed boundary mesh. The method we proposed is initially based on DeWall algorithm[3]. This algorithm can generate a Delaunay tetrahedrization for a
given set of points, producing a valid (but may be low-quality) FEM mesh for the points’ convex hull (CH). The method we propose extends it with some advancing front-like operations in order to meet the additional constrains, and as a side effect it simplifies the dealing with degenerated elements or ambiguous configurations (something that the original approach cannot solve in practice). The result is a highly decoupled method that can fit both shared and local memory architecture.

![Figure 1: DeWall’s scheme: Given a virtual Delaunay complaint boundary (initial front B1), the algorithm will choose a plane (W1), build all elements with edges intersected by the plane updating the front, until there are no more edges over the plane. The space is then partitioned into an already meshed subregions and two unconnected unmeshed ones (B2 and B3 will be the new Delaunay boundaries). The same algorithm can be applied again in the resulting unmeshed subregions (choosing W2 and W3 planes).](image)

### 2.1 DeWall algorithm

DeWall algorithm is based on a triangulation method described by McLain for the interpolation of random data[7] and the Divide and Conquer paradigm (D&C). The input is an unordered set of points, and the generated output is the Delaunay triangulation or tetrahedrization for the convex-hull of the given points. The idea is to choose a line (for 2D problems) or plane (for 3D problems) and build all elements intersected by such line/plane. We’ll call this line/plane the divisor. As a result, the divisor will partition the overall domain in two unconnected subdomains. After generating the elements along the line, two or more unconnected subdomains remain unmeshed and can be solved individually without the need for communication. The same algorithm is applied recursively (Figure 1) until each subdomain consists on a single element. That’s one of the keys for the paralellization strategy. When the process starts, there’s only one thread working, but after the first iteration, two threads can work in parallel, and ideally four in the third, eight in the fourth, and generalizing $2^n$ in the $n$-th iteration. The number of busy processors grows exponentially with main iterations, so idle processors won’t remain idle for a long time when the process starts. Also, due to the uniqueness of the Delaunay triangulation (in general position settings), each subdomain generates a set of elements whose boundary will fit the elements in the original plane(s) that defined such subdomain, avoiding the need of extra processing in a merging stage.

The method McLain propose for building the Delaunay triangulation consists on choos-
ing a point to build a triangle with a given base edge known to be part of the Delaunay triangulation. Given an initial base edge, the process start by selecting the first point as potential (candidate) third point for the new triangle, and calculate the center of circle through the three points. The important measure to consider is the signed distance from that center point to the base edge (Figure 2). The point which minimized that distance is the point that should be used to build a Delaunay compliant triangle. All following points are processed testing whether they lay inside or outside that circle. A point inside the circle would have a lower distance, and will become the new candidate point (and it’s circle the new testing circle). A point outside the circle can be dismissed. If a point lays just in the limit of the circle, it is an ambiguous case for a Delaunay triangulation (for instance, the four vertex from a perfect square). In such cases, a rule for solving the tie such that the same result is guaranteed if the triangle where being build with one of the other two edges as base edge must be applied. There are many simple theoretical solutions for this problem that work perfectly in the context of an infinite precision arithmetic, but will fail in a computer implementation due to the numerical error implied in its floating-point computations. Big care must be taken when implementing solutions for this to conserve all subproblems generated by the D&C strategy uncoupled and their results compatible. This problem will be addressed in section 2.2.

![Figure 2](image)

**Figure 2**: Left: McLain’s criteria for building a new triangle given a base edge \((P1 - P2)\) from current front. \(D\) is the measured distance for a feasible point \(F\). Right: Four feasible points \((F1...F4)\) with their respective circumcenters \((C1...C4)\). \(C4\) is the winner according to the signed minimum distance criteria (it has a negative distance).

After a triangle is built, the two new edges can be used as base edge for new triangles. The parallel method will choose only the edge that’s being intersected by the line. In a 3D version, the base edge becomes a base face and the circle becomes a sphere, but the overall idea is exactly the same. The only remaining problem is how to find an initial edge/face intercepting the divisor to start with, so this is done just once, and any \(O(n)\)
algorithm can be used without having a sensible impact on overall performance. There’s is a simple method in the original article but it is not important here because we will have an imposed boundary to pick from.

The final key element for this method is to use an spatial sorting data structure to avoid testing every point when building a new triangle with a given base edge. If points are tested ordered by increasing distance from the edge or from current candidate’s circum-center, massive discard can be done reducing the initial $O(n^2)$ complexity to a sub-quadratic one, with expected linear times. An evaluation of the performance and a comparison with other Delaunay Triangulation algorithms can be found in [7].

2.2 Imposed boundary

The main difficulty in fitting an imposed boundary mesh, is the impossibility to build it with a Delaunay Tetrahedrization when the given boundary is not Delaunay-compliant. Any valid mesh constrained to such a boundary will include some non-Delaunay elements. If we break the Delaunay property in the already presented algorithm, we can no longer rely on that generated meshes from subproblems will merge together without requiring some local edition to fix incompatible interfaces. Subproblems can then became coupled. Furthermore, even with a valid Delaunay boundary mesh, the presence of ambiguous configurations can lead to the same sort of problems.

Many approaches where tested for solving this problem. The main idea is to start the first main iteration with the imposed boundary as the initial front, build elements with some additional tests when selecting a candidate point for an edge/face, and set the new generated edges/faces as part of the front for following iterations and later for the next subproblems. The additional tests for each candidate point consists on several intersection tests to ensure that the potentially new element the point will generate with the current base edge will be compatible with the current front. We call this points feasible points. So a feasible point for a given base edge is a point that generates a triangle that doesn’t intersect the current front. Solving this problem for the 2D version is relatively easy. The winner point for a given base edge will be the best one according to the Delaunay criteria but only considering feasible points. By defining imposed boundaries for every subproblem from the components of the resulting front, the problem of ambiguous Delaunay configurations is automatically solved (any solution chosen for an ambiguous configuration is imposed to future subproblems). In 2D, any imposed boundary with any set of given points inside can be solved with a valid triangulation without the need to add, move nor remove points. In the 3D version of the algorithm, this can lead to unsolvable subproblems. One way to overcome that problem is to include in the process a rollback mechanism (a way to undo some elements and retry), and a method to avoid regenerating the same elements again. More details about the proposed rollback mechanism will follow in next section.

Considering that non-Delaunay elements will be usually needed near the original boundary and for ambiguous situation where numerical errors can invalidate the test, in most
cases Delaunay elements will be generated. In order to speed up the process, we propose a two-pass strategy to find out the fourth point for a given base face. In the first pass we ignore the intersections’ tests, generating a potential element as fast as the original DeWall algorithm does. Then we test this element for intersections with the current front, and only if any intersection exists, we run a second pass where we repeat the whole search for a fourth point for the base face, this time enabling all tests. The number of cases where the second pass is required is very low compared to the total number of generated elements in most cases. Only when many ambiguous points settings (more than 4 co-spherical points) are found inside the domain this second stage will be frequently run.

Finally, it is important to state that relaxing some numerical thresholds inside the intersection tests, these tests can detect some low-quality elements (slivers and caps) very common in Delaunay 3D meshes. So these tests can be used to prevent the grid generator from generating these sort of elements without any extra cost. This feature can be enabled in every job as long as it haven’t done too many rollback operations. If a job seems to be stalled placing and removing elements without successfully filling its domain, the algorithm will shrink the thresholds in order to finish the mesh at the cost of generating a few low-quality elements. However, this improvement will increase significantly the number of rollbacks operations required, and this can be very time-consuming as will be shown later.

3 Implementation

This section focuses on the 3D version of the proposed algorithm. We call the problem of building all elements trough the dividing plane for a given domain a job. A job starts with an imposed boundary that can be thought as an advancing front. Some face from this front that crosses the divisor plane is selected and a new tetrahedron built upon that base face is added to the mesh. After that, the base face won’t be in the front anymore, but the other three element’s faces might be (for each face if it was not already in the front, otherwise this indicates that the front closes there and so the face should also be removed). The same process is repeated until the job ends when there are no more faces touching the divisor plane. As a result, the front would be generally subdivided into two unconnected subproblems: one to the left of the dividing plane, and one to the right. Actually, this subproblems could be void or, in concave domains, could be composed by more than one unconnected components. The set of nodes from the original problem is also partitioned by the plane in two or three subsets: one to left, one to the right, and may be one additional group with nodes that lays right on the plane (this nodes won’t be required by neither subproblem).

As stated before, a rollback mechanism will be required to ensure algorithm generality. This mechanism will be needed on relatively low number of iterations. Such addition, when not necessary, do not generate a noticeable overhead neither in memory consumption nor in running time compared to the original DeWall algorithm.
3.1 Nodes spatial sorting

There are two methods to avoid testing every node when building a new element for a given base face. The first option is to associate nodes to jobs and to test only the current job’s nodes. Moving nodes container to jobs, or reordering them inside the nodes vector is not an option because this would invalidate all faces and tetrahedron’s references. Knowing that a node cannot belong to more than one job at the same time, we proposed building an additional array of references to nodes (that will be part of the mesh structure), and reordering that array in order to keep all nodes from a given job consecutive. The array starts with all nodes (for the initial job), and after the first divisor plane is meshed, all nodes to the left of that plane can be relocated at the beginning of the array, and all nodes to the right at the end. The same procedure can be applied recursively for every job. A job only needs to know the limits of its range in that array. The second option is using some spatial sorting data structure to sort nodes and use the sphere property described before to massively discard them. For instance, if the spatial sorting data structure is an octree, nodes from the octree’s cells that don’t intersect the current sphere in some point of the algorithm won’t need to be tested. Timing measures showed that a very simple structure such as a regular grid outperforms more complex ones such as bucketed-octrees. A regular grid has constant query times, is easier to traverse by neighbor cells, and is more cache-friendly. The ratio of nodes per cell may be unbalanced for meshes with variable node density but testing proved this not to be a real problem and lead to almost linear times for the overall process. Both methods are actually combined to achieve the best performance. If the number of nodes in a subproblem domain is low the first method is used, otherwise the second one is. The threshold to choose when to switch can be empirically fitted for a given hardware.

When testing nodes by the order given by the grid (or any other spatial sorting data structure, the same idea will apply), we need some simple and fast method to avoid testing a grid cell twice as part of the same job for the same base face. The simpler way is to use a binary flag per cell that will mark which cells where already visited for a given base face, but this means that we need to reset a lot of flags for every element built. To avoid this problem we use a very simple trick: the flag is not just a binary variable, but an integer instead. We start with all flags set to 0, and an extra integer next_flag with value 1. In the process of building an element, an element’s flag will be considered set if its value is the same as next_flag, and unset otherwise. When an element is finally built, instead of resetting all flags, we just increase next_flag, and that will mean that all elements’ flags will now be automatically unset. There could be a problem if next_flag overflows, but this will only happen once every $2^{32}$ elements (it is reasonable to reset all flags in that particular case). In order to allow parallel execution without duplicating the grid and without requiring extra synchronization, each cell will actually have an array of a few integers, one per thread. Notice that this implementation seems to be likely to produce the problem known in parallel programming as false-sharing[1], but comparing times with
a different implementation to avoid this (aligning in memory the per-cell arrays, or using per-thread flags arrays), showed it is not noticeable in this case.

3.2 Rollback mechanism

A rollback operation is called when there’s no feasible node for a given face from the current front. In that case, the element that generated the base face, and/or one of the elements that invalidate candidate nodes (that intersect a potential new tetrahedron) have to be removed, and there must also be some way to avoid rebuilding it/them again later (to avoid an infinite loop). In order to restore the integrity in the set of faces that compose the advancing front, every face in the front has a reference to the tetrahedral element that generated it (or a null reference if its part of the boundary) and every tetrahedral element keeps track of its neighbors. So when the algorithm cannot continue for a given front, the faces’ references are used to select the elements that can be undone, and the elements’ references will be use to set the new faces after being removed.

To avoid infinite loops, every state of the front known as impossible will be stored in another data structure. So there is an extra requirement for a node to be considered feasible for a given base face. It is that the front that this node would generate if selected should not be present in set of known impossible fronts. This set is an additional data structure (based on a set or a hash table implementation) that should be associated to the job. To minimize the need for this additional verification, when the algorithm cannot create a new element for some face in the front, it will try to advance as much as possible with all other faces in the same front, before saving the front as impossible. Doing so, the saved front will be smaller, and so will the remaining unmeshed region inside that front. In most cases where a rollback operation is needed, this situation appears in the last few elements, and require very few iterations (elements deletions).

Summing up, there is a very clear situation that signals the need for a rollback: there’s a base face in the current front for which no candidate point can be found. But it can be very expensive for the algorithm to detect this situation, given that the absence of any feasible point in such scenery prevents the grid from discarding any cell based on an initial candidate points’ sphere.

It is also very important to set a criteria to choose which or what elements to remove when rollback is needed. This decision will have a direct impact in overall running times, because a wrong choice leads to another impossible front (whose detection can be, as stated before, very expensive). Several deterministic criteria where implemented and tested, but for all of them we could relatively easily find a very bad worst case scenario (in terms of running time). Finally, the best criteria seems to be a two stage one: first delete all non-Delaunay elements and retry, and then if the problem persists delete single random ones. Elements are flagged as Delaunay or Non-Delaunay on creation, knowing if when creating them from a base face, there was o wasn’t a better point in terms of Delaunay criteria but discarded by the intersections’ tests.

Since impossible fronts are usually very small (composed by less than 10 faces), it
would be very fast approach if the detection of the impossible front where also fast. It’s not always the case. An impossible front in some of the first jobs is very expensive to detect due to the big number of points to test. The front advances from outside to inside, from the job’s boundary to the center of the job’s domain. So, when it reaches its first impossible setting, it is near the center of its domain. Only if the domain is small, the problem can be near the imposed boundary (either for the whole mesh, or for that single job) So, it is less likely to find an impossible front in the first jobs, than it is in the last ones, when jobs’ domains are very small. This makes the penalty time taken to detect those situation not so long in most cases.

3.3 Intersection tests

Testing the feasibility of a given point for a base triangle in this 3D meshing algorithm is one of the most expensive parts. For every tested candidate point (that now implies the point was already tested according to Delaunay criteria) we must build the faces and edges that the potential new element would generate, and try to intersect them with the faces and edges from the current front. This front can be arbitrarily big (it is initially the imposed boundary), so an efficient data structure for selecting a reduced set of elements from that mesh to be tested is required. Many advancing-front mesh generators make use of Alternate Digital Trees (ADT)[2][9]. This data structure is designed to query axis aligned bounding box (AABB) intersections efficiently. The basic idea is to build a tree with 6D points for storing 3D AABB. The 6D correspond to the lower and upper bounds of a 3D AABB. This data could be stored in any kind of sorted three supporting multidimensional keys. In particular, for a fixed set of elements a range tree[4] will exhibit very good search times. But since every new element modifies the current boundary mesh, a structure that allows fast element insertion and deletion is required. For this purpose, and ADT fits better. It operates like a binary search tree, but on every depth level the comparison between elements for searches is done only with the coordinate of one of the six dimensions.

The proposed implementation for the described method requires one ADT per job, implemented as a bucketed binary-search tree. It must be also noticed that the ADT and the current front will contain the same set of faces, so all face elements insertions and deletions have to keep these two structures’ contents synchronized. In order to do so the ADT actually contains references to elements in the lists that represents the current front, implemented as iterator-like proxy classes, with enough information to delete an element from its list.

3.4 Threads synchronization

A shared memory approach can make use of a single mesh object where every job inserts elements in, and a single spatial sorting data structure for nodes queried by every job. As a consequence, the implementations have to use mutexes, semaphores and other
typical mechanisms for avoiding many potential race conditions and deadlocks. Also, some scalar properties that cannot be shared between parallel jobs are replaced by vectors in a shared memory architecture (a single id from the serial implementation might need to be replaced by a vector of ids, one per thread). Our reference implementation only makes use of a mutex to synchronize write access to the elements list, and a condition variable to implement a thread-safe jobs queue. Every other required data is either local to a job (example: the ADT), by design warranted not to be accessed simultaneously by two jobs (examples: nodes ids, grid cells tags), or read-only (example: nodes coordinates).

In a local memory strategy, every job has its own copy of the mesh and related data structures. The need to duplicate the data and the need to merge the results have obvious overheads. But the time needed to duplicate the input data has no big impact in the overall process when this duplication is performed in a single machine and so no network communication is required (a local memory strategy running on a shared memory architecture). The time for the merging stage is in this situation negligible since merging here only means collecting all elements in a single container, but no tests or extra operations are needed. The most significant extra time in these stages is the time needed to build the spatial sorting data structures (such as the grid). But there are two positive consequences from this approach: jobs will work with partial meshes, so will work with smaller inputs; and no thread synchronization mechanism at all is required inside a job. Smaller inputs means more locality and smaller data structures that will lead to faster query times and a better use of the different cache levels in the computer architecture. This can lead to super linear speedups in a 2D implementation, where data structures and intersection tests are far more simple[8]. However, this is not the situation in the 3D implementation, where data structures are less cache-friendly, and profiling showed that intersections tests related operations are the most time consuming ones (and these are cpu-bounded operations).

In both strategies, load balancing can be easily done. It means, to prevent one thread from being busy with a huge job while other threads are idle waiting for new ones. Every job is smaller than the previous job that generated it, so the jobs queue will hold jobs approximately ordered by size, and thus can choose to solve bigger jobs first and let small ones for ensuring a good load balance at the end of the process.

4 Results

This section introduces results from a shared-memory implementation, tested on a 1st generation I7 quad-core architecture. A preliminary analysis with some specific profiling tools indicates that this process is cpu-bound for low number of threads, as it is the case in today’s high-end desktop PCs. If the number of available real processors increases, then the local-memory strategy discussed before could be more suitable even for shared memory architectures (furthermore considering NUMA). It also shows a very good load balance from the point in the execution when there are enough jobs for all threads to be working, on. Even if this point is reached very quickly and almost without time penalty
for these PCs, further work is needed in order to seize the power of idle processors in this early stage for more complex architectures.

Figure 3 shows how the proposed algorithm scales linearly in time with the size of the mesh. Times do not consider the variation to avoid low-quality elements. This variation introduces a big deviation in measured times and still needs further work to produce stable results in terms of timed performance. Finally, figure 4 shows how the algorithm scales when adding more processors. Parallel efficiency varies from 90% to 64.7% for several measured scenarios with 1 to 5 processors (note that the test machine only has 4 real processors, the fifth one is virtually achieved via Intel’s Hyper Threading technology, that expose 4 extra virtual cores to the software and is expected to produce from 15% to 30% performance gains in several applications[5]).
5 Conclusions

DeWall algorithm has been successfully extended to fit situations where a fixed boundary mesh have to be met. Some algorithms and spatial sorting data structures common in advancing front mesh generators where adapted to fit our needs and the most significant implementation guidelines where discussed. Experimental results have showed that the algorithm complexity is still competitive, and parallel efficiency and scalability are already very good in shared memory architectures, even when some noticeable drawbacks have been identified. We where able to successfully apply it to non general-position settings, and it can easily avoid some types of low-quality elements very common in Delaunay grids, but with some noticeable time-penalty. The final algorithm is also suitable for local memory architectures without major changes, but further work is still needed to reach the full potential of this technique in the mentioned directions.

REFERENCES


ASSESSMENT OF EDGE-BASED FINITE ELEMENT TECHNIQUE FOR GEOPHYSICAL ELECTROMAGNETIC PROBLEMS: EFFICIENCY, ACCURACY AND RELIABILITY

OCTAVIO CASTILLO¹, JOSEP DE LA PUENTE¹, VLADIMIR PUZYREV¹ AND JOSÉ MARÍA CELA¹

¹Barcelona Supercomputing Center
Nexus I - Campus Nord UPC - Barcelona - Spain
contact: octavio.castillo@bsc.es

Key words: Finite Element, Geophysics, Electromagnetism and Edge Elements

Abstract. In Finite Element Methods for solving electromagnetic field problems, the use of Edge Elements has become very popular. In fact, Edge Elements are often said to be a cure to many difficulties that are encountered (particularly eliminating spurious solutions) and are claimed to yield accurate results [9, 16, 17]. We will shortly describe the mathematical formulation of linear edge elements and we go through the particular issues related to the implementation of these elements in order to solve geophysical electromagnetic problems. In particular, we describe a simple, flexible and parallel Fortran 90 implementation for Edge Elements. The code is based on an abstract data structure, which allows to use different kinds of solvers with little effort. The result is an implementation that allows users to specify Edge-based Finite Element variational forms of $H(\text{curl})$. Finally, we also show the performance of the code in terms of efficiency, accuracy and reliability, which will shape our future line of work in order to solve more complex problems.

1 INTRODUCTION

The Sobolev spaces $H(\text{div})$ and $H(\text{curl})$ play an important role in many applications of Edge-based Finite Element methods (Edge Elements) to partial differential equations (PDE). Examples include second-order elliptic PDE’s, Maxwell’s Equations for electromagnetism, and the linear elasticity equations, among others. Edge-based Finite Element methods may provide advantages over standard $H^1$ Finite Element discretizations in terms of improved flexibility, stability and robustness. However, implementing $H(\text{div})$ and $H(\text{curl})$ methods requires additional code complexity for constructing basis functions
and evaluating variational forms, which together with its relative novelty helps explaining their relative scarcity in practice.

The most important difference between the Nodal Finite Element (Lagrange FE) and the Edge-based Finite Element method (Nédélec Elements) is that the basis functions are not defined on the nodes of 2D triangular and 3D tetrahedral meshes, but on edges and faces, respectively [14]. Edge Elements provide only partial continuity over element boundaries: continuity of the normal vector component for $H(\text{div})$ problems and continuity of the tangential vector component for $H(\text{curl})$ problems.

The method of FE applied to $H(\text{div})$ and $H(\text{curl})$ problems and its implementation has been well documented [1, 16, 17]. Users can find many software codes such as NGSOLVE [3] or FEniCS [12], which are written in object oriented languages allowing for higher order elements defined on elements with curved boundaries. Such codes are well suited for high complexity computations and also provide a certain flexibility via user interface. However, if some features are not available, it is usually difficult to understand the source code and modify it. We believe that our Fortran 90 code is more suitable for students and researchers who wish to become familiar with Edge Elements and prefer to make their own modifications. As our target application is exploration geophysics, our computational solution considers the lowest order linear Edge Elements defined on 2D triangles for $H(\text{curl})$ problems only. There is a large bibliography dedicated to constructing nodal elements, but only few publications related to Edge Elements. Our implementation takes may practical ideas from literature [1, 6, 16, 17].

The paper is divided as follows: Section 2 shortly describes the theory associated to linear edge elements without going into details. In Section 3 we go through the particular constructions related to the implementation of these elements. Strong emphasis is placed on aspects not easily found in the literature, such as the choice of orientation of geometric entities. Section 4 describes the software stack, including the most important functions such as the interface that allows to use different kinds of solvers with little effort. In particular, we use the BLAS library [2]. Section 5 shows the performance of the code in terms of efficiency, accuracy and reliability. Finally, we make some concluding remarks in Section 6.

2 LINEAR NÉDÉLEC ELEMENTS FORMULATION

When using nodal elements (Lagrange elements) in electromagnetism, spurious solutions can occur [9, 10, 17]. In this section we summarize some basic facts about the Sobolev space $H(\text{curl})$ and we discuss conforming finite elements associated with them. The reader is referred to [16] for a more thorough analysis of $H(\text{curl})$.

We denote by $\Omega$ an open, bounded and connected Lipschitz domain in $\mathbb{R}^d$, where $d \in \{2, 3\}$ denotes the space dimension. The rotation of a vector valued function $w : \Omega \to \mathbb{R}^d$ is defined as:
\[ \nabla \times w = \begin{pmatrix} \partial_2 w_3 - \partial_3 w_2, \\ \partial_3 w_1 - \partial_1 w_3, \\ \partial_1 w_2 - \partial_2 w_1, \end{pmatrix} \] (1)

Following the approach of [16, 17], we consider two types of rotation operators in 2D, the vector operator \textbf{curl} and the scalar operator \textbf{curl}:

\[ \textbf{curl} f = \begin{pmatrix} \partial_2 f \\ -\partial_1 f \end{pmatrix} \] (2)

\[ \text{curl} w = \partial_1 w_2 - \partial_2 w_1 \] (3)

In literature, the operator \textbf{curl} is frequently called the co-gradient. The operators give rise to the standard Sobolev spaces:

\[ H(\nabla \times \Omega) = \begin{cases} \{ v \in L^2(\Omega \mathbb{R}^3) \mid \nabla \times v \in L^2(\Omega \mathbb{R}^3) \} & \text{if } d = 3 \\ \{ v \in L^2(\Omega \mathbb{R}^2) \mid \nabla \times v \in L^2(\Omega) \} & \text{if } d = 2 \end{cases} \] (4)

where \( L^2 \) denotes the space of the square Lebesgue integrable functions. Assuming that \( \Omega \) is discretized by a triangular (2D) or a tetrahedral (3D) mesh \( \mathcal{M} \), Nédélec Elements represent basis functions in \( H(\text{curl}, \mathcal{M}) \) spaces. Figure 1 shows the numbering of the degrees of freedom of the linear Edge elements.

On the other hand, we denote the global edge basis functions by \( n_i \), and by \( x = (x_1, x_2, x_3)^T \) the spatial variable in \( \Omega \). Similarly, we define the reference basis functions and the spatial variable simply by adding the hat \( \hat{\cdot} \), for instance, \( \hat{x} \) denotes the spatial variable in the reference element. The reference basis functions of the Nédélec Elements are as follows [1, 14, 16, 17]:

\[ \text{2D case} : \begin{align*} n_1(\hat{x}) &= \begin{pmatrix} -\hat{x}_2 \\ \hat{x}_1 \end{pmatrix}, \\ n_2(\hat{x}) &= \begin{pmatrix} -\hat{x}_2 \\ \hat{x}_1 - 1 \end{pmatrix}, \\ n_3(\hat{x}) &= \begin{pmatrix} 1 - \hat{x}_2 \\ \hat{x}_1 \end{pmatrix} \end{align*} \] (5)

\[ \text{3D case} : \begin{align*} n_1(\hat{x}) &= \begin{pmatrix} 1 - \hat{x}_3 - \hat{x}_2 \\ \hat{x}_1 \\ \hat{x}_1 \end{pmatrix}, \\ n_2(\hat{x}) &= \begin{pmatrix} \hat{x}_2 \\ \hat{x}_1 \\ \hat{x}_2 \end{pmatrix}, \\ n_3(\hat{x}) &= \begin{pmatrix} \hat{x}_3 \\ \hat{x}_3 \\ 1 - \hat{x}_2 - \hat{x}_1 \end{pmatrix}, \\ n_4(\hat{x}) &= \begin{pmatrix} -\hat{x}_2 \\ \hat{x}_1 \\ 0 \end{pmatrix}, \\ n_5(\hat{x}) &= \begin{pmatrix} 0 \\ -\hat{x}_3 \\ \hat{x}_2 \end{pmatrix}, \\ n_6(\hat{x}) &= \begin{pmatrix} \hat{x}_3 \\ 0 \\ -\hat{x}_1 \end{pmatrix} \end{align*} \] (6)

In the following, \( F_K \) denotes the affine element mapping \( F_K := B_K \hat{x} + b_k \) from the reference element to an element \( K \) in the mesh. In order to preserve the tangential
continuity of the reference basis functions, we need to use the so-called Piola mappings. We refer the reader to [4, 13] for a more thorough treatment. In our case, the values are mapped as follows [1, 14, 16, 17]:

\[ n(x) = B_K^{-T} n(F_K^{-1}(x)) \]  

Finally, the rotation is mapped differently depending on the dimension [1, 16, 17]:

2D case: \[ n(x) = \frac{1}{\det B_k} \text{curl} n(F_K^{-1}(x)) \]  

3D case: \[ n(x) = \frac{1}{\det B_k} B_K \text{curl} n(F_K^{-1}(x)) \]

3 PARTICULARITIES OF EDGE ELEMENTS

In this section we shortly describe three particularities to construct an Edge Elements solution: Piola mapping in order to guarantee global continuity of \( H(\text{curl}) \) problems, numbering scheme and edges direction.

3.1 Piola mapping

First, it follows from Stokes theorem that in order for piecewise \( H(\text{curl}) \) vector fields to belong to in \( H(\text{curl}) \) globally, the traces of tangential components over patch interfaces must be continuous. In order to do that, one must consider the covariant Piola mapping which is defined by:

\[ F^{\text{curl}}(\Phi) = J^{-T} \Phi \circ F^{-1} \]  

Figure 2 depicts the vector field \( \Phi \) between two triangles using the covariant Piola mapping which preserves tangential traces of vector fields [16].

We refer the reader to [4, 16, 17] for a more thorough treatment.
3.2 Numbering strategy

In order to guarantee global continuity with Piola mapping, special care has to be taken with regard to the numbering strategy. Our implementation is based on UFC (Unified Form-assembly Code), which is a unified framework for finite element assembly. More precisely, it defines a fixed interface to communicate low level routines (functions) for evaluating and assembling FE variational forms [12].

In short, the numbering strategy works as follows. A global index is assigned to each node of the mesh $M$ (consisting of triangles or tetrahedra). If an edge adjoins two nodes $n_i$ and $n_j$, we define the direction of the edge as going from node $n_i$ to node $n_j$ if $i < j$. This gives a unique orientation of each edge. The same philosophy is used locally to determine the directions of the local edges on each element. Thus, if an edge connects the second and the third node of a tetrahedron, then the direction is from the second to the third node.

A similar numbering scheme is employed for faces, the key point now is to require that the nodes of each element are always ordered based on their global indices. More details may be found in [9, 13].

3.3 Edges direction

Since the degrees of freedom are integrals over edges or faces, we need to know how they are oriented. To do that, we need to include what is the positive orientation for an edge or a face in the mesh $M$. Then, we proceed as follows. First, for every element we have an affine mapping from the reference element to the element in the mesh. Second, the reference element has a certain orientation for the edges and faces. Thus, if the orientation of an edge, or the orientation of a face is mapped in the same direction as positive direction we had agreed upon, we assign $+1$ for this edge/face. Otherwise we assign $-1$. These directions are depicted in Figure 3 [1, 16, 17].

In 3D, we need to again know which edge unit tangential vectors to use, hence, the calculation of the orientations is identical to the 2D case.
4 SOFTWARE STACK

In this section we describe the software stack of our solution including the most important data structures. An outline on the primary groups of functions in our implementation is given in Figure 4. A more detailed explanation is the following:

1. Mesh. This module stores geometric and topological properties of a mesh: how are the elements connected and where are their nodes. The properties and data of meshes are almost always queried through loops over all elements, possibly querying all faces of each element as well (3D). Our implementation is able to read as input nodal-based meshes.

2. Edge-FE. Describe the properties of a Edge-based Finite Element space as defined on the reference element. This includes, for example, individual shape functions at edges on the reference element.

3. Quadrature. As with FE, quadrature objects are defined on the reference element. Includes the location of quadrature points on the unit element, and the weights of quadrature points thereon.

4. Initializer. This module is the confluence of meshes and Edge-FE, in other words, Edge-FE describes how many degrees of freedom it needs per node or edge, and Initializer module allocates this space so that node or edge of the mesh has the correct number of them. It also gives them a global numbering.

5. Mapping. Computes matrix and right hand side (RHS) entries or other quantities on each element in the mesh, using the shape functions of a Edge-FE shape functions and quadrature points defined by a quadrature rule. To do that, it is necessary to map (Affine mapping) the shape functions, quadrature points and quadrature
weights from the unit element to each element in the mesh. This is directly done by Mapping module. It also provides support to compute signs for each edge in 2D or faces in 3D.

6. EdgeFE_values. This module takes the Edge-FE and evaluate its shape functions and their values at the points defined by a quadrature rule when mapped to the real element.

7. Assembly. This module assembles the system matrix and RHS of the linear system. We will determine the solution of our problem from this linear system. To this end, we have subroutines that store and manage in an efficient way the entries of matrices and vectors.

8. Solver interface. In order to find the solution of a linear system, one needs linear solvers. In FE applications, they are frequently iterative, but sometimes one may also want to use direct solvers. Since our implementation takes advantages of external solvers such as BLAS Library or PETSC to name a few, we need the Linear solver module, which allows communication with the external solvers.

9. Output. Once one has obtained a solution of a Edge-FE problem on a given mesh, one will to postprocess it using a visualization program. Our software doesn’t do the visualization by itself, but rather generates output files with the final results [12]. It also gives timing values in order to evaluate the performance.
10. Common modules. Finally, our implementation has three fundamental services: parallel service, memory service and I/O service. In other words, common modules is a toolbox that provide a variety of independent procedures to be called by other modules or subroutines.

About data structures, our implementation is based on ideas of [1, 5, 9]. We consider triangles in 2D and tetrahedra in 3D and denote by \( \#\varpi \) the number of elements in the set of \( \varpi \), and by \( \mathcal{N}, \mathcal{E}, \mathcal{F} \) and \( \mathcal{T} \) the sets of Nodes, Edges, Faces and Elements, respectively. Table 1 describes the elemental matrices that are needed in order to implement Edge Elements. All structures are based on Column-Major Order approach.

<table>
<thead>
<tr>
<th>Name</th>
<th>Dimensions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>nodes2coord</td>
<td>2/3 x #\mathcal{N}</td>
<td>Nodes defined by their 2/3 coordinates in 2D/3D</td>
</tr>
<tr>
<td>edges2nodes</td>
<td>2 x #\mathcal{E}</td>
<td>Edges defined by their 2 nodes in 2D/3D</td>
</tr>
<tr>
<td>faces2nodes</td>
<td>3 x #\mathcal{F}</td>
<td>Faces defined by their 3 nodes in 3D</td>
</tr>
</tbody>
</table>

With the previous matrices available, we can express every element by the list of its nodes, edges or faces as states Table 2, note that faces \( \mathcal{F} \) exist only in 3D case.

<table>
<thead>
<tr>
<th>Name</th>
<th>Dimensions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>elems2nodes</td>
<td>3/4 x #\mathcal{T}</td>
<td>Elements by their 3/4 nodes in 2D/3D</td>
</tr>
<tr>
<td>elems2edges</td>
<td>3/6 x #\mathcal{T}</td>
<td>Elements by their 3/6 edges in 2D/3D</td>
</tr>
<tr>
<td>elems2faces</td>
<td>4 x #\mathcal{T}</td>
<td>Elements by their 4 faces in 3D</td>
</tr>
</tbody>
</table>

We recall that we use first-order elements. Hence, in 2D, if one uses the unit triangle as the reference configuration, Nédélec Elements have a degree of freedom related to each of the three edges. In 3D case, if one uses the unit tetrahedron as reference configuration, Nédélec Elements will have a degree of freedom related to each of the six edges. Hence, the structures elems2nodes, elems2edges and elems2faces define the numbering of global degrees of freedom for a given mesh. Figure 5 provides a simple example of the mesh data in 2D.

5 PERFORMANCE ISSUES

In order to test the performance of our solution, we focus in assembly time, solving time, and convergence order. Our tests are based on ideas from [1], where the authors solve...
the eddy-current problem using Nédélec Elements. The experiments were performed on a simple node of the Marenoutrum supercomputer with two 8 - core Intel Xeon processors E5 - 2670 at 2.6 GHz.

Uniform refinement in 2D results in 4 times more elements. In this work, the BLAS library is compiled and linked with our code in order to use the SGESV subroutine as solver [8]. Table 3 depicts the performance of our code with 16 OpenMP threads in terms of assembly time and solver time. Table 3 also includes the mesh spacing (h), $L^2$ error and the convergence order which is plotted also in Figure 6.

**Table 3:** Summary of results for the 2D case. Number of elements (T), number of edges (E), assembly time (seconds), solver time (seconds), mesh spacing (h), $L^2$ error and convergence order.

<table>
<thead>
<tr>
<th>#T</th>
<th>#E</th>
<th>Assembly time</th>
<th>Solver time</th>
<th>h</th>
<th>$L^2$</th>
<th>$O_{L^2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>128</td>
<td>208</td>
<td>$7.66 \cdot 10^{-3}$</td>
<td>$6.76 \cdot 10^{-3}$</td>
<td>$3.3 \cdot 10^{-1}$</td>
<td>$3.969 \cdot 10^{-1}$</td>
<td>-</td>
</tr>
<tr>
<td>512</td>
<td>800</td>
<td>$6.75 \cdot 10^{-3}$</td>
<td>$7.72 \cdot 10^{-3}$</td>
<td>$2.0 \cdot 10^{-1}$</td>
<td>$9.831 \cdot 10^{-2}$</td>
<td>2.016</td>
</tr>
<tr>
<td>2,048</td>
<td>3,136</td>
<td>$2.22 \cdot 10^{-2}$</td>
<td>$4.47 \cdot 10^{-2}$</td>
<td>$1.1 \cdot 10^{-1}$</td>
<td>$2.430 \cdot 10^{-2}$</td>
<td>2.017</td>
</tr>
<tr>
<td>8,192</td>
<td>12,416</td>
<td>$1.82 \cdot 10^{-1}$</td>
<td>$2.08 \cdot 10^{-1}$</td>
<td>$5.8 \cdot 10^{-2}$</td>
<td>$6.001 \cdot 10^{-3}$</td>
<td>1.999</td>
</tr>
<tr>
<td>32,768</td>
<td>49,408</td>
<td>$4.27 \cdot 10^{-1}$</td>
<td>$6.68 \cdot 10^{-1}$</td>
<td>$3.0 \cdot 10^{-2}$</td>
<td>$1.501 \cdot 10^{-3}$</td>
<td>2.004</td>
</tr>
<tr>
<td>131,072</td>
<td>197,120</td>
<td>$1.51 \cdot 10^{0}$</td>
<td>$2.27 \cdot 10^{0}$</td>
<td>$1.5 \cdot 10^{-2}$</td>
<td>$3.741 \cdot 10^{-4}$</td>
<td>2.002</td>
</tr>
<tr>
<td>524,288</td>
<td>787,456</td>
<td>$5.83 \cdot 10^{0}$</td>
<td>$6.71 \cdot 10^{0}$</td>
<td>$7.8 \cdot 10^{-3}$</td>
<td>$9.340 \cdot 10^{-5}$</td>
<td>2.002</td>
</tr>
<tr>
<td>2,097,152</td>
<td>3,147,776</td>
<td>$21.32 \cdot 10^{0}$</td>
<td>$24.7 \cdot 10^{0}$</td>
<td>$3.9 \cdot 10^{-3}$</td>
<td>$2.333 \cdot 10^{-5}$</td>
<td>2.001</td>
</tr>
</tbody>
</table>

Figure 7 depicts the discrete solution of eddy-current problem in a mesh with $T = 32,768$ and $E = 49,408$. 
6 CONCLUSIONS

The relative scarcity of Edge-based Finite Elements in practical use may be attributed to their higher theoretical and implementational cost. Indeed, more care and effort are required to implement them: basis functions, Piola mapping, edge directions and numbering strategy. However, as demonstrated in this work, the implementation of Nédélec Elements for $H(\text{curl})$ conforming may be automated. Particularly, the additional challenges in the assembly process can be viewed as not essentially different from those encountered in other approaches such as higher order Lagrange elements.

The first version of our framework, based on a simple mathematical approach [1], is able to resolve $H(\text{curl})$-conforming problems in 2D. The described software stack relies on a flexible implementation which allows a general point of view. The efficiency and accuracy of the code is evaluated through a convergence test, assembly time and solver time, with a particular emphasis on the performance when the number of elements and degrees of freedom grows.

In our future work we will implement the 3D case, and we will also address issues that are related to the optimization of the code, i.e. solution techniques and in particular efficient assembly algorithms and preconditioning techniques that reduce the global system size and allow for a significant speed-up of the linear system solver. In this context, we will also investigate on domain decomposition techniques and parallelization in the spirit of [10, 11, 15]. Another natural extension of the presented framework is to address...
Figure 7: Discrete solution of eddy-current problem: curl(above), x-component(bottom-left) and y-component(bottom-right).

Inhomogeneous domains and adding dipole sources.

We believe that our Fortran 90 code is suitable for students and researchers who wish to become familiar with Edge Elements and prefer to have their own implementation.

REFERENCES


CONSTRUCTION OF SCHOOL TIMETABLE WITH EMPHASIS IN MINIMIZATION OF TWINNED LESSONS AND WINDOWS

MÁRCIA INÊS SCHABARUM MIKUSKA¹, CARLOS EURICO GALVÃO ROSA², ANDERSON ROGES TEIXEIRA GÓES¹ AND SÉRGIO SCHEER¹

¹Federal University of Paraná, Graduate Program on Numerical Methods in Engineering
P.O Box 19011, Zip Code 81531-980, Curitiba-PR, Brazil
mmikuska@ig.com.br, artgoes@ufpr.br, scheer@ufpr.br

²Federal University of Paraná, Advanced Campus Jandaia do Sul
426, Doutor João Maximiano street, Zip Code 86900-000, Jandaia do Sul-PR, Brazil
cegalvao@ufpr.br

Key words: Timetabling Problem, optimization, metaheuristics.

Abstract. The timetabling problem is, as best as possible, allocate teachers at their respective classes, to meet with their students, so that it complies with the amount of weekly lessons of discipline. Many basic education institutions still do this work manually, since commercial softwares to solve this problem are considered high cost, especially for public schools, and softwares available on the web sometimes do not give good solutions. Thus, is demanded a lot of time in the pursuit of meeting the preference of requests from teachers, maybe creating windows, twinned lessons or isolated lessons. In this context, this research seeks to resolve problems such as those mentioned using a metaheuristics and compare the quality of the solution obtained with the use of exact method since metaheuristics admit possible to determine a feasible solution even not getting the optimal solution to the problem. Moreover, the mathematical modeling problem will be penalized the solutions having the undesirable characteristics, increasing the likelihood that the model itself select the solutions with better performance.

1 INTRODUCTION

During the planning of a school year, usually at the beginning of the year of the pedagogical week, each educational institution seeks to define what will be the time that teachers of subjects meet with their students. On this occasion, teachers inform their schedules limitations and conditions for the lessons, such as the presence or prohibition of twinned lessons and it is up to those responsible for teaching coordination to find a solution in which meet these availabilities given the amount of required lessons in the school curriculum and also avoiding windows.

There are various softwares, among paid and free, used by some institutions. Most of the time a presented solution can eventually cause problems like unwanted excess windows or a grouping of many times to the same course for a class on the same day, seriously affecting
learning. On the other hand, obtains still very common a solution to this problem by hand, mainly in small public and private schools. Often this timetable is drawn up in the pedagogical planning period, involving teachers and coordination and is not always easy to get a consensus on the quality of the solution.

2 PROBLEM DESCRIPTION

The timetabling problem studies began in the 1960s by Appebly [1] and Gotlieb [2]. In 1995, due to interest from researchers, was the first specific international conference on Practice and Theory on Automated Timetabling - PATAT, which occurs every two years to the present day.

The school timetabling problem is an NP-hard or NP-complete problem due to lots of variables and constraints created in the modeling, increasing the time of computational resolution.

As analyzed by Góes, Costa & Steiner [3], a lot of work done in this area of research involves undergrad programs, where the weekly timetable has not to be necessarily complete, (may have idle intervals between lessons), which results in a flexibility not assigned to Brazilian's Basic Education institutions like Elementary and High School. According Pillay [4], the problem is described as compact because is without free or idle periods. For these institutions, this problem becomes more complex, because the classes are taught at all times and a class change of a teacher will result in a change in another teacher.

On [3] are the following considerations for this type of problem in Basic Education Institutions:

- A predetermined number of hours per day (in this paper, daily workload) and per week (weekly workload) to the class must be satisfied;
- The weekly workload of a subject to each class must be satisfied;
- A lesson must not have more than one teacher in the same class
- A class must not have more than one teacher at the same time;
- A teacher must not minister a lesson to more than one class at the same time;

If some obtained solution does not comply with these situations above, its implementation is infeasible and will be obligatorily discarded. In addition to these, is needed to verify if the proposed schedule is accordance with the teacher's availability. For teachers with an employment bond with only one institution, common in institutions dominated by Exclusive Dedication schemes, the availability of the teacher arises often his preference and implementation of other activities inherent to the position as lesson preparation, counseling students and research. For these it is possible to assign a schedule of orientation and hours of planning, in here called "activity hour", if predicted by the school. On the other hand, for teachers who teach in more than one institution, the compliance with the time availability is critical to be able to meet its commitments.

There may be other situations that do not affect the viability of the solution, but its quality. In these cases the solution presented can be performed, but hampers the learning process. Examples of these situations: excess consecutive lessons of the same course in the same class and the same day, called twinned lessons; idle spaces of time between lessons where teachers
are no lessons, here called \textit{windows} and the number of days with only one lesson in every period, here called \textit{isolated lessons}.

The twinned lessons' occurrence can be defined by a preference of the teacher (GÓES, 2005). However, a long explanation of the same subject in continuous time is generally tiring both for the student as for the professional. The windows can be unpleasant for teaching in their extreme cases, when the teacher is allocated in the first and last lesson, making infeasible take other compromise or be absent between these lessons. In addition, for private institutions, the window is counted as working time and generates a cost increase with the teacher. At long last, the occurrence of windows in some days may cause the existence of isolated lessons on other days of the week.

2.1 The Real Problem - Willy Janz School

For realization of the techniques that will be implemented, it was decided to apply them in the timetable generating of the Willy Janz School in 2014. For realization of the techniques that will be implemented, it was decided to apply them in the timetable generating of the Willy Janz School. It is a private school belonging to the Institute Mennonite Brethren Biblical Seminary (ISBIM Portuguese acronym) with Christian principles offering from kindergarten, preschool and Primary School. The timetable is set for the morning shift, which focus junior high school classes (In Brazil, "Ensino Fundamental II"), namely the sixth to the ninth grade. Due the existing demand at that time and physical structure available, the school offered two classes of the 6th grade (6th A and 6th B), and one class for each seventh, eighth and ninth grade, as shown in Table 1. According to National Education Guidelines and Framework Law (LDB Portuguese acronym) [5] the disciplines of the Portuguese Language, Mathematics, Science, History, Geography, Physical Education and Arts belong to the Common National Base, while Music, English and Biblical Studies make up the diverse part.

\begin{table}[h!]
\centering
\begin{tabular}{lcccccc}
\hline
 & 6th A & 6th B & 7th & 8th & 9th \\
\hline
Portuguese Language & 6 & 6 & 5 & 5 & 5 \\
Mathematics & 5 & 5 & 6 & 6 & 6 \\
Science & 3 & 3 & 3 & 3 & 3 \\
History & 2 & 2 & 2 & 2 & 2 \\
Geography & 2 & 2 & 2 & 2 & 2 \\
Physical Education & 2 & 2 & 2 & 2 & 2 \\
Arts & 1 & 1 & 1 & 1 & 1 \\
English & 2 & 2 & 2 & 2 & 2 \\
Biblical Studies & 1 & 1 & 1 & 1 & 1 \\
Music & 1 & 1 & 1 & 1 & 1 \\
\hline
\end{tabular}
\caption{Number of lessons (weekly) per subject / class}
\end{table}

In table 1 above, it is possible observing that there is a difference in the number of Portuguese Language and Mathematics lessons between classes.

The workload of each teacher in each class is shown in Table 2 below.
Table 2: Teachers’ weekly workload per class

<table>
<thead>
<tr>
<th></th>
<th>6th A</th>
<th>6th B</th>
<th>7th</th>
<th>8th</th>
<th>9th</th>
<th>Weekly workload</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1 – Mathematics</td>
<td>5</td>
<td>5</td>
<td>---</td>
<td>---</td>
<td>6</td>
<td>16</td>
</tr>
<tr>
<td>P2 – Mathematics</td>
<td>---</td>
<td>---</td>
<td>6</td>
<td>6</td>
<td>---</td>
<td>12</td>
</tr>
<tr>
<td>P3 – Portuguese Language</td>
<td>6</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td>12</td>
</tr>
<tr>
<td>P4 – Portuguese Language</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>P5 – Arts</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>P6 – Geography</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>P7 – History</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>P8 – English</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>P9 – Biblical Studies</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>P9 – Physical Education</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>P10 – Music</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>P11 – Science</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
<td>15</td>
</tr>
</tbody>
</table>

Whereas are taught 27 hours per week in Portuguese Language and 28 hours in Mathematics adding all classes, it is not possible for a single teacher takes every class, which does not happen to other disciplines. There is the same identifier to the disciplines of Biblical Studies and Physical Education because these subjects are taught by the same teacher. The difficulty in drawing up the schedule of a private school is that most teachers have employment contracts with more than one institution, also teaching in other schools. The availability of days to teach in each school is very inflexible. It is more preferable to schedule the maximum of "complete days", or without windows. Another occurrence in the institution refers to about 2/5 of the teachers of the institution who attend graduate programs and, therefore, also require specific days to go watch their lessons. There may be changes in these availabilities during a year, according to the graduate program calendar, whether quarterly or half-yearly.

3 APPROACHES PROPOSALS IN THIS RESEARCH

“A general timetabling problem includes scheduling a certain number of events (exams, courses, meetings, etc) into a limited number of time periods, while satisfying as many of the required constrains as possible” [6]. The restrictions for this type of problem can be classified as:

- Hard constraints: cannot be violated under any circumstances. Violate those constraints will affect the solutions’ viability. For example, a teacher cannot be in two different classes at the same time.
- Soft constraints are desirable, but not essential for resolution. Those constraints are affecting the solution' quality. For example, the existence of windows.
3.1 Mathematical Modeling - Exact Method

The essence of the solution is set up a teacher, linked to a class, is (or is not) teaching on a particular day and time, given certain constraints. As for each teacher, class, day and time have only "Yes" and "No" answer options; it is a binary integer linear programming problem.

Thus, a software was developed in Visual Basic that generates this model with its constraints and objective function for the exact solution method by LINGO Software (Language for Interactive General Optimizer), version 13.0.

The objective function lists the $x_{pda}$ variables associated with the presence of the teacher $p$ in the class $t$ day $d$ lesson $a$. The $p_{pda}$ values, attributed to the availability of the teacher $p$ on day $d$ lesson $a$, are coefficients of objective function. This is due to the fact that there are teachers with time restrictions on certain days, such as having available only the last or first lessons. If the teacher has not availability in a whole day, it is considered that is not available in any of the lessons for that day. With this, the objective function is:

$$Z = \text{Max} \sum_{p} \sum_{t} \sum_{d} \sum_{a} p_{pda} \cdot x_{pda} - \Phi$$

where $p \in \{1, 2, \ldots, \text{quantity of teachers}\}; \ t \in \{1, 2, \ldots, \text{quantity of classes}\}; \ d \in \{1, 2, \ldots, \text{days per week}\}; \ a \in \{1, 2, \ldots, \text{lessons per day}\}$ and $\Phi$ is the function that records the penalties obtained in the non fulfillment of soft constraints.

The main constraints of the problem, based on [3], modeled follow below:

$$\sum_{p} x_{pda} = 1, \ \forall t, \forall d, \forall a$$

$$\sum_{t} x_{pda} \leq 1, \ \forall p, \forall d, \forall a$$

$$\sum_{d} \sum_{a} x_{pda} = WW_{pt}, \ \forall p, \forall t$$

$$\sum_{a} x_{pda} = DW_{pdt}, \ \forall p, \forall t, \forall d$$

with $WW_{pt}$ weekly workload of teacher $p$ class $t$ and with $DW_{pdt}$ daily workload of teacher $p$ class $t$, day $d$. For the disciplines of Portuguese Language and Mathematics, was assigned $DW = 3$ for every day and classes. For the other disciplines are used maximum $DW = 2$. This is due to the case of teachers who need to 3 lessons in the same class, without being twinned.

The constraints (2) ensure that in each class there is only one teacher per schedule. Assurance that each teacher will be a maximum of one time per class is the fulfillment of the constraints (3). Constraints (4) and (5) ensure that will be complied with respectively the weekly and daily workloads. Compliance with these constraints ensures the viability of the proposed solution.

Considering only this step the problem presents a considerable number of variables and constraints. For the case study are 11 teachers who teach five classes. How are five lessons per day and five days a week, the problem have $11 \times 5 \times 5 \times 5 = 1375$ variables. About constraints set out above are 125 constraints in (2), 275 constraints in (3), 55 constraints in (4)
and 275 constraints in (5), totaling 730 restrictions. A valid solution must contain exactly one value "1" for each tuple (the "class-day-lesson" set) and all other values equal to zero to comply with the requirement that each class has one, and only one, teacher in all his days and lessons. It means a search space of \( \binom{1375}{125} \approx 3 \times 10^{180} \) possible solutions. The computational time to check this search space is very large. A supercomputer making around \( 33,68 \times 10^{15} \) calculations per second would take almost \( 3 \times 10^{153} \) millennia [7].

Soft constraints are evaluated in individuals considering the existence of windows, the amount of twinned and isolated lessons, if undesirables. This value is set in the objective function as a penalty, and as a maximizing problem, the trend is that this value will be minimized.

### 3.2 Metaheuristic Based on Genetic Algorithms

The initial proposal for the research in question is based in [3], which is the phase that is the study.

The evolutionary computing can be classified as one solution strategy designed to solve generic problems and operate in non-linear and non-stationary areas. It is accepted the fact that it is not guaranteed total efficiency in obtaining the solution. However, you can get a good approximation to the optimal solution, considering that the computation time increases at a lower rate than exponential [8]. Some situations that occur in nature were studied and compiled for computation - Artificial Intelligence. Examples of these are genetic algorithms, local search and particle swarm.

The Artificial Intelligence with Genetic Algorithms (GA) researches was initiated by Holland in 1975 [9]. In that, were formalized and mathematically explained the processes of adaptation in natural systems and developed artificial systems computationally simulated, keeping the features found in the natural order in natural systems. Were kept the crossover operations, which is the exchange of information between chromosomes of selected individuals, and mutation, which is a modification of information in a chromosome.

Although not initially been proposed to work with optimization problems, GA has been much applied to this branch of research. Some common terms in GA are in Table 3.

<table>
<thead>
<tr>
<th>Natural</th>
<th>GA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>Set of solutions to the problem</td>
</tr>
<tr>
<td>Individual</td>
<td>A solution to the problem</td>
</tr>
<tr>
<td>Chromosome</td>
<td>String which comprises the solution</td>
</tr>
<tr>
<td>Gene</td>
<td>Feature</td>
</tr>
<tr>
<td>Allele</td>
<td>Feature value</td>
</tr>
</tbody>
</table>

Hamawaki [10] advocates the use of GA because being very efficient to search for optimal solutions, or approximately optimal, with applications in many problems. Due to the
limitations found in traditional methods exact, GA have the ability to identify and explore environmental factors and converge to the optimal solutions in global levels.

The GA’s functionality begins with the initialization of the population with fitness calculations, selection, reproduction and mutation. According to Goldberg [11], in natural systems, one or more chromosomes are combined to form the "genetic recipe" for the construction and operation of an organism. Parents have their chromosomes combined to generate offspring. In GA, the application of the described operations in several individuals creates a new population, or new generation, which should have a better approximation of the solution, compared with the previous population. For the initial population, values will be randomly assigned to the gene of each chromosome, which can be represented in binary, integer or real types. The fitness of an individual in the population is measured by means of an objective function: The fitness calculation. Often, the fitness of the best individual and a limitation to the number of generations is used as criteria to stop the algorithm.

![Figure 1: Structure and function of GA](image)

In this paper, “individual” is a possible weekly timetable of each class. An individual is considered viable when comply with the hard constraints. As described in the modeling, soft constraints are evaluated on improving the solution, directly influencing at the individual's fitness value.

![Figure 2: Example of an individual](image)

The subjects are represented by their teachers in order to address situations in which a single teacher is responsible for more than one subject. These identifiers are shown in Table 2. The set of individuals that satisfy all classes is considered “population”.

To create a new generation will have to use genetic operators. On [3] only mutation was used as a genetic operator to obtain more adapted individuals. So the authors entitle his work as “GA based metaheuristic”. The crossover procedure generates a great amount of unviable
individuals. With mutation, however, it is possible to recover lost information from previous generations.

The initial generation was made as follows:

1. Pick up a random teacher in a list of possible teachers.
2. For each class, in order, check if the professor teaches to this class. In positive case, analyze the teacher's availability. If available, assign lessons to the teacher in this day as many as possible. Applies this procedure to each day and each class, avoiding schedule clashes.
3. If the weekly workload of this teacher has not been completed, apply the same procedure, but allowing schedule clashes.
4. Remove up the teacher of the list of possible teachers to pick.
5. Repeat the previous steps until exhaust the teachers list.

At this early generation, it is not guaranteed adaptability of individuals, even their viability, because there is a possibility of constraints violation. These situations will be corrected by changes defined by an improvement heuristic.

Created a generation, the next step is to evaluate it, calculating its fitness value. The first value to be considered will be find the teacher's availability. For each lesson given with availability add a positive value to the fitness. Lessons allocated to unwanted moments will be penalized by adding negative values to fitness.

A second evaluation will be done comparing the number of working days with the required amount of working days. For example, a teacher who has 10 hours weekly workload will work at least two days, and the two complete days. If it is allocated for 3 days, maybe there are windows in this timetable. Therefore, each additional day worked must be penalized, favoring solutions that have a minimum number of days for each teacher.

The evaluation of twinned lessons existence will be done by analyzing the occurrence of these classes, according was told by the teacher and considering the weekly workload and characteristics of subjects. Example, even most of students wish, it is impossible considering Physical Education as twinned lessons, as well as three consecutive Math lessons when a teacher has three Math lessons in his daily workload for that class.

Finished the evaluation of population, it is necessary to define the mutation operator that will be used. It will work to correct violations of the constraints as hours excess in the daily working per class, more days working than the minimally necessary and the schedule clashes.

[3] draw up a list of existing errors on individuals:

1. List the teachers with schedule clashes;
2. List the teachers who have exceeded daily workload per class;
3. List the teachers with more work days than necessary;
4. If no teacher has been listed in the previous steps, there are no errors to be treated and the scan finishes (stopping criterion). Otherwise, the scan continues.

5. Choose one of the teachers listed. This may be the teacher with the least amount of errors or a random choice.

6. List up the timetable problems of the chosen teacher. Pick up one of the listed errors and this error will be handled by a specific routine.

7. The previous steps are repeated until comply with stopping criteria.

Specific improvement routines involve changes of classes and they were developed for each detected violation type. In carrying out the mutations, it is possible that some errors are created, like schedules clashing in a teacher when carrying out the mutation with another teacher, some allocations creating undesirable twinned lessons, among others. It is the GA that allows us to infer that these errors will be minimized because their existence negatively influences the individual's fitness, making it less likely to survive in the following generations.

The individual's survival in a population at the GA context follows some selection methods to verify if the individual must join the new generation. One method is the bi-classist selection in which, according to Carvalho & Yamakami [12], in each generation is preserved P% of the best individuals and (100 – P)% of the worst individuals in the population. Another widely used method is the elitist model, where only the P% of the best individuals remaining in the population. As the population is a possible solution to the problem of this work, was used the following method of selection:

1. Created a new population, evaluate its fitness;

2. If 70% of individuals have a better fitness than the previous population, in other words, in 70% of the classes there was an improvement in the appointment of teachers, according constraints under which the problem was subjected; it is accepted throughout the population and algorithm continuous generating a new population. Otherwise, follows the present analysis;

3. If the new population is between 50% and 70% better fitness than the previous population, it is possible to accept this new population, according some probability and, after acceptance, will be proceeding to the next generation. If was not complied the percentages or the probabilistic method has rejected the population follows the analysis;

4. Declines all generated population and a new solution is generated from the previous, returning to the beginning of this analysis.

These steps are repeated until a predefined number of generations. This way of solution has great inspiration in the manual resolution, but has its execution managed by the GA, seeking the increase in solution quality.
4 CONCLUSION

The processing done with the linear programming problem by LINGO® software on computer with Intel® Core™ i3-3217 processor, 1.8 GHz CPU, 4.0 Gb of RAM, obtained an optimal solution in 1 second, which shows that the problem is feasible. The use of the proposed metaheuristic intends to improve the quality of this solution through any relaxation of soft constraints, which do not make infeasible the solution.

The timetabling problem belongs to school context and, despite the existence of automated solutions; these are not used by ignorance or a possible cost. The comparison between grids generated is difficult to be done by a percentage analysis, because a solution can present as a single, but fatal, failure. For example, violation of the availability of a teacher condemns all a solution. At this point it has shown the usefulness of the management process done by the GA based algorithm: the violations affect the fitness of the solution, decreasing the chances of it to be taken into account. Continuous improvement in solution quality leads to the expectation of going towards the optimal, since given the vastness of the space of possible solutions becomes impossible to examine one by one all solution.

To continue this work should be done more analysis with the results achieved, improving the methods used in order to get better and better solutions, seeking to implement other metaheuristics for solving it.

There is also the intention to provide the results of this research for use in educational institutions in solving their timetable problems.

ACKNOWLEDGMENTS

The authors would like to thank UFPR, the PPGMNE and professors for the valuable guidance and opportunities. We extend our thanks to Willy Janz School, represented by its principal Prof. Roberto Renner, by the data provided and support in the implementation of this project.

REFERENCES


PERFORMANCE ASSESSMENT OF ABSORBING CONDITIONS FOR THE REVERSE TIME-HARMONIC MIGRATION

Hélène Barucq$^{1}$, Morgane Bergot$^{2}$, Juliette Chabassier$^{1}$ and Julien Diaz$^{1}$

$^{1}$ Inria Team-Project Magique-3D, University of Pau, France
helene.barucq@inria.fr, juliette.chabassier@inria.fr

$^{2}$Institut Camille Jordan, University Claude Bernard Lyon 1, France
bergot@math.univ-lyon1.fr

Key words: Wave Propagation, Numerical Simulation, Absorbing Boundary Condition, Reverse Time Migration

Abstract. We propose an imaging technique based on the solution of harmonic wave equations. Its principle is inspired by the Reverse Time Migration which is widely used for seismic imaging. The numerical method involves Radiation Boundary Conditions and we assess the impact of these conditions on the capability of the imaging method to recover an obstacle hidden inside a fluid.

1 INTRODUCTION

Waves are widely used for providing images of the subsurface. By making use of their reflection capabilities, it is possible to create a map identifying each change of medium. This map deserves a particular attention because it contains relevant information on the material characteristics of the propagation domain without requiring any extraction process. Reverse Time Migration (RTM) is an imaging technique based on the solution of time-dependent wave equations set inside bounded regions given as the domain to be explored. RTM provides a way for putting time information recorded by some receivers back to the piece of subsurface of interest. RTM is well-known to be a robust seismic imaging technique but it is computationally intensive. It is indeed based on the solution of a large number of wave equations that distinguish themselves by the source term and the construction of the image requires a pre-processing step which often reaches the limitations of the memory. There is thus a need in optimizing the computational costs of RTM. Several tries have been achieved by developing new numerical schemes for High-Performance Computing (HPC) of wave equations [4, 6]. It turns out that numerical costs can be reduced but this improvement is limited by the validity domain of the discretization. Reducing numerical costs actually gives advantage to explicit time-schemes.
which work under stability conditions. Obeying the stability condition then freezes the reduction of computational costs. Moreover, the construction of the image is based upon a discrete imaging operator defined on a cartesian grid which needs to be fine to capture heterogenities correctly. As a consequence, RTM is not easy to put in action when targeting realistic problems because it generates a huge computational burden shared between computational time and memory storage. We then propose to assess the performance of RTM in the harmonic domain by suiting the RTM algorithm to harmonic waves for providing images more quickly. Our approach is different from what is done for Full Waveform Inversion since we do not minimize a cost function. By this way, we may avoid issues of local minima while taking advantage of the robustness of harmonic equations in the sense that we do not have to perform any time discretization.

This paper provides preliminary results for motivating our solution methodology. We address a simpler problem which consists in recovering the shape of an obstacle. The computational domain is limited by an artificial boundary and the numerical waves are then obtained by solving harmonic wave equations coupled with specific boundary conditions on the artificial boundaries which are generally referred as Absorbing Boundary Conditions (ABC) [12] in the time-domain or Radiation Boundary Conditions (RBC) [9] in the harmonic domain. It is well-known that the accuracy of the solution to Helmholtz equation depends on the RBC that is employed. This is why, in addition to the validation of a harmonic-wave seismic imaging technique, we investigate how the RBC can impact on the quality of the image.

2 RADIATION BOUNDARY CONDITIONS

In this section, we introduce the boundary conditions that we use for solving the Helmholtz equation. Their derivation is explained in detail in an internal report [5] dealing with high-order artificial boundary conditions both for the harmonic and time-dependent wave equation. We recall that the Helmholtz equation reads as:

\[ \frac{\omega^2}{c^2} u + \Delta u = 0, \]

where \( \omega \) denotes the pulsation and \( c \) is the propagation velocity. The equation is set in a two-dimensional bounded domain \( \Omega \) with an external boundary \( \Sigma \) representing the artificial boundary. Let \( \kappa(s) \) be the curvature of the boundary \( \Sigma \), \( s \) denoting the curvilinear coordinate. We then consider the first order family of RBCs which are given by:

\[ k = \frac{\omega}{c}, \quad (\partial_n u + i k u) + \left(\gamma + \frac{k}{4}\right) u + \left(\gamma - \frac{k}{4}\right) \frac{\partial_n u}{ik} = 0, \]

where \( \gamma \) denotes a parameter and \( n \) stands for the normal vector on \( \Sigma \) outwardly directed to \( \Omega \). It has be proven in [7] that if \( \gamma \) is larger than \( \kappa(s)/4 \), then the boundary value
The problem of Helmholtz equation coupled with this RBC is well-posed. When \( \gamma \) is equal to \( \kappa(s)/4 \), we get the most widely used RBC for the Helmholtz problem:

\[
\partial_r u + ik u + \frac{\kappa}{2} u = 0.
\] (3)

The condition is often referred to as “curvature-RBC” or “C-RBC” ([14, 13, 9]). The conditions (2) have been derived by approximating a non-reflecting boundary condition under the assumption that the waves impinge the boundary with an incidence angle close to the normal one. It is interesting to note that RBCs can also be constructed by assuming the frequency \( \omega \) is high. In that case, the lowest-order approximation gives rise to:

\[
k = \frac{\omega}{c}, \quad \partial_r u + ik u + \left( \gamma + \frac{\kappa}{4} \right) u = 0.
\] (4)

Then it is worth to note that the specific value \( \gamma(s) = \frac{\kappa(s)}{4} \) also leads to the “C-RBC” previously obtained in [14, 13, 9]. By performing the next order of approximation, we then have:

\[
k = \frac{\omega}{c}, \quad \left[ 1 + \frac{\gamma - \kappa}{4ik} \right] \partial_n u + \left[ ik + \gamma + \frac{\kappa}{4} \right] u - \frac{1}{2ik} \Delta \Sigma u = 0.
\] (5)

When \( \gamma(s) = \frac{\kappa(s)}{4} \), we get a well-known second-order RBC:

\[
\partial_n u +iku + \frac{\kappa(s)}{2} u - \frac{1}{2ik} \Delta \Sigma u = 0.
\] (6)

3 NUMERICAL IMAGING BASED ON HARMONIC WAVE SOLUTIONS

The Reverse Time Migration [11] is based on the cross-correlation of waves. The numerical waves are divided into two groups: waves generated by sources and numerical waves corresponding to the back-propagation of reflected waves. The cross-correlation of these two groups gives then the picture of reflectors, that is each point of the propagation domain where reflections occur. We follow the same algorithm but by solving harmonic wave equations. We then have a first group of numerical waves generated by source points \( x_j, j = 1, ..., N \), which are dispatched regularly around a penetrable obstacle inside the computational domain (see Fig. 1). The waves are reflected by the boundary of the obstacle to be detected and the reflected waves are recorded by sensors. Then the
recordings provide a set of complex data which are conjugated and used as sources in the Helmholtz problem. By this way, we solve the conjugate set of equations which amounts to back-propagate the reflected waves. We then end up with solving two Helmholtz problems with multiple right-hand sides given as the different sources. The procedure to follow is then:

**Propagation of sources**: solve

\[
\begin{cases}
\frac{\omega^2}{c^2} u_{S_j} + \Delta u_{S_j} = S_j, \ j = 1, \ldots, N, \\
\text{RBC on } \Sigma.
\end{cases}
\]  

where \( S_j, \ j = 1, \ldots, N \) denotes a point source at a given point \( x_j \): \( S_j = \delta_{x_j} \). For each \( j \), record \( u_{S_j}(x_i), i = 1, \ldots, N \).

**Back-propagation of reflected waves**: solve

\[
\begin{cases}
\frac{\omega^2}{c^2} u_{R_j} + \Delta u_{R_j} = \sum_{i=1}^{N} \bar{u}_{S_j}(x_i)(x_i) \delta_{x_i}, \ j = 1, \ldots, N, \\
\text{RBC on } \Sigma.
\end{cases}
\]  

![Figure 1: Position of the sources and of the obstacle](image)

When real data are not available, the first experiment must be performed twice to provide a set of data. In that case, we construct the set of data by using another numerical method. By this way, we avoid inverse crime.
Once the two problems have been solved, we need to construct the image by applying an imaging condition. Herein, we propose the imaging operator $I$ defined at each point $x$ of the numerical grid by:

$$I(x) = \sum_{j=1}^{N} \Re(u_{S_{j}}(x) \ast u_{R_{j}}(x)).$$

4 NUMERICAL RESULTS

The numerical experiments are performed by using an Interior Penalty Discontinuous Galerkin approximation [1] that is implemented in the code Hou10ni, using curved elements to discretize the external boundary and the obstacle. We refer to [8] for more details on the implementation of Hou10ni. The computational domain is a circle of radius 2km, the obstacle has the shape presented in Fig 1. Inside the obstacle, we set $c = 2200$ m/s and outside the obstacle we set $c = 1500$ m/s. We use 100 sources placed on a circle of radius 1.9 km (very close from the boundary) and we set the frequency $\omega = 50$ Hz. We have used finite elements of degree 4. The global linear system is solved with Mumps, using the multi-right hand side feature. This feature is another advantage of harmonic RTM compared with time-dependent RTM, which requires to solve one wave equation for each source.

For the time being, we have consider the family of RBCs (4) for various values of $\gamma$. It turns out that all the images have the same quality and that we are always able to identify the obstacle. That is why we set only a picture (see Fig.4) which illustrates the capability of the numerical method to reconstruct the shape of the obstacle.

We are thus able to identify the obstacle by applying the extension of RTM in the harmonic regime. We have considered various RBCs and they all provide the same solution. More numerical experiments are needed to conclude because we have considered a simple case where the artificial boundary is a circle. In that particular case, it has been demonstrated (see for instance [7]) that all the conditions (4) are equivalent. High-order conditions should thus be considered with different shapes of artificial boundary.

5 Concluding remark: about injecting grazing wave modeling

It has been illustrated in [7] that the lack of performance of classical RBCs may come from the fact that the grazing waves are not controlled by the usual RBCs. Following page 140 of [7], it is possible to enforce the control of grazing waves by modifying classical RBCs. For instance, the modified C-RBC reads as:

$$\left[1 + \frac{c - \kappa/2}{ik}\right] \partial_{n}u + \left[ik + c + \frac{\kappa c}{2ik}\right] u - \frac{1}{ik} \Delta_{\Sigma}u = 0. \quad (9)$$

For example, let us consider the case where the obstacle is a sound-soft elongated ellipse. We solve the first problem with one source and we apply the modified C-RBC on the artificial boundary $\Sigma$. Its shape is in keeping with the shape of the obstacle. The parameters
of the obstacle are 0.5 for the small axis and 5 for the large axis. The reference solution is computed by solving the Helmholtz equation coupled with a non-reflecting boundary condition [10] set on a boundary which must be put far enough from the obstacle. Basically, it is necessary to have several wavelengths between the obstacle and the artificial boundary. However, the numerical solution obtained with the modified RBC is computed by using an artificial boundary which is set very close to the obstacle (see figure 3). The real part of the reference solution $u_{ref}$ is displayed for a frequency equal to 1 (see figure 3). The table 5 shows that adding the modeling of grazing waves improves the relative L2-norm error $err$ defined by

$$err = \sqrt{\frac{\int_{\Omega} |u - u_{ref}|^2}{\int_{\Omega} |u_{ref}|^2}}.$$ 

This is particularly clear when the frequency increases.

We thus observe that there is an advantage in including grazing waves inside the external boundary condition. In the context of seismic imaging involving harmonic waves, we now aim at investigating if by using this kind of condition, it might be possible to put the artificial boundary closer to the sources. This could contribute to reduce the computational costs. According to a previous work on On-Surface-Radiation Conditions [3], we believe that modified RBCs including grazing wave modeling might help. Anyway, modified RBCs should improve the quality of the image since it has been shown in [7] that the
solution of the Helmholtz equation is really more accurate. However, these conditions involve higher order differential operators which means that their numerical handling will increase the computational burden. This point deserves a particular attention because this could be an inconvenient when considering three-dimensional problems.

REFERENCES


DROPLET IN MICRO-CHANNELS: A NUMERICAL APPROACH USING AN ADAPTIVE TWO PHASE FLOW SOLVER

Jose-Maria Fullana, Yue Ling, Stéphane Popinet, Christophe Josserand

1 Sorbonne Universités,
UPMC Univ Paris 06, UMR 7190, Institut Jean Le Rond d’Alembert,
F-75005, Paris, France

2 CNRS, UMR 7190,
Institut Jean Le Rond d’Alembert,
F-75005, Paris, France 4 Place Jussieu, Paris, France
jose.fullana@upmc.fr

Key words: Computational Mechanics, microfluids, two-phase flows

Abstract. We propose a numerical approach to study the mechanics of a flowing bubble in a constraint micro channel. Using an open source two phase flow solver (Gerris, gfs.sourceforge.net) we compute solutions of the bubble dynamics (i.e. shape and terminal velocity) induced by the interaction between the bubble movement, the Laplace pressure variation, and the lubrication film near the channel wall. Quantitative and qualitative results are presented and compared against both theory and experimental data for small Capillary numbers. We discuss the technical issues of explicit integration methods on small Capillary numbers computations, and the possibility of adding Van der Walls forces to give a more precise picture of the Droplet-based microfluidic problem.

1 INTRODUCTION

Droplet-based microfluidics is a very promising tool for performing biochemical or chemical assays. Droplets are unit systems of controlled volume and content, within which mixing can be easily achieved. Several physical phenomena (mechanics, thermocapillarity, solutocapillarity, thermomechanics) either in cumulative or compensative ways appears when we develop microfluidic setups. It is of prime importance to characterize, under controlled experimental conditions, within which range each contribution is the dominant phenomena regarding element migration. Rationalizing these various effects would have important consequences for lab-on-a-chips, and numerical studies are an interesting way to understand each contribution separately.
In microfluidic setups we often have measurements of shape deformations, bubble or drop velocities, and pressures or flow rates at the input/exit conditions, but the knowledge local values of these variables are not easily available, because principally of the small length scales involved in the system. Numerical approaches are then an interesting way for acceding to small length dynamical fields. The validity of numerical approaches requires the validation and the confrontation against theories and experimental data. Theoretical results exist in few academic microfluidic configurations like bubble in cylinders or plates, and it is then necessary to be able to compare positively these academic configurations before extend the prediction to more complex situations which are actually common in microfluidic devices.

In this communication we present an open source two phase flow tool (gfs.sourceforge.net) for computing solutions of the bubble dynamics (i.e. shape and terminal velocity) in microfluidic channel and; numerical results are presented and compared, quantitatively and qualitatively, against both theory and experimental data for small Capillary numbers. We discuss the possibility of adding Van der Walls forces to give a more precise picture of the Droplet-based microfluidic problem.

2 EQUATIONS AND NUMERICAL SCHEME

We use the incompressible, two-dimensional variable-density, NavierStokes equations with surface tension which can be written

$$\rho \left( \frac{\partial U}{\partial t} + U \nabla U \right) = -\nabla p + \mu \nabla^2 U + \sigma \kappa \delta \hat{n}$$

$$\nabla U = 0$$

with $U = (u, v)$ the fluid velocity, $\rho = \rho(x, t)$ the fluid density, $\mu = \mu(x, t)$ the dynamic viscosity. The Dirac distribution function $\delta$ expresses the fact that the surface tension term is concentrated on the interface; $\sigma$ is the surface tension coefficient, $\kappa$ and $\hat{n}$ the curvature and normal to the interface.

For two-phase flows we introduce the volume fraction $c(x, t)$ of the first fluid and define the density and viscosity as a function of $c$, i.e. $\rho = \rho(c(x, t))$ and $\mu = \mu(c(x, t))$. The advection equation for the density can then be replaced with an equivalent advection equation for the volume fraction

$$\partial_t c + \nabla (Uc) = 0$$

The Navier-Stokes equations are solved using a finite volume approach based into a projection method. The numerical problem is solved using the open-source package Gerris [6]. A staggered in time discretisation of the volume-fraction/density and pressure leads to the a formally second-order accurate time discretisation. The interface between the different fluids are tracked and followed using a VOF (Volume of Fluid) method. The spatial discretisation is done using a quad-tree square cells which give a very important
flexibility allowing dynamical costless grid refinement into user-defined regions. Finally a powerful discretisation scheme was developed to capture accurately the surface tension term. More information can be found in reference [1]

3 BUBBLE IN A CHANNEL

The studied configuration is presented in Figure 1: a bubble is pushed into a microchannel of width $H$ by a mean flow velocity $U_f$. The typical length of the bubble is larger than the width $H$, the bubble is then constrained by the channel. In the stationary regime the velocity bubble is $U_b$. Using the width $H$ as characteristic length and the mean flow velocity $U_f$ as characteristic velocity the dimensionless Navier-Stokes is then written

$$\frac{\partial U}{\partial t} + U \nabla U = -\nabla p + \frac{1}{R_e} \nabla^2 U + \frac{1}{R_e C_a} \kappa \delta \hat{n}$$

(2)

where all dynamical variables are dimensionless and we define the Reynolds and Capillary numbers as

$$R_e = \frac{\rho U H}{\mu}$$

(3)

$$C_a = \frac{\mu U_f}{\sigma}$$

(4)

Over this communication we do not discriminate densities and viscosities for liquid and bubble as long as in the subsequent computations we fix the density and the viscosity ratios between the fluid and the bubble to one, then $\rho_f/\rho_b = 1$ and $\mu_f/\mu_b = 1$. Bretheron [2] studied theoretically and experimentally the dynamic of a bubble on a cylindrical configuration. In the limit of small capillarity number based on the bubble velocity, it is shown that the ratio between the gap of the thin film of lubrication $h$ (between the wall and the bubble) and the typical height $H$ of the channel as well as the ratio between the bubble and mean fluid velocity scale both as $C_a^{2/3}$:

$$\frac{h}{H} \sim \frac{U_b}{U_f} \sim C_a^{2/3}$$

(5)
the proportionality constant depends, at least, on the geometrical configuration (planar, squared or cylindrical) and on the viscosity ratio. By inspecting the $h/H$ relation and its dependency into the capillary number $C_a$, it appears that the grid refinement plays an important role if we want to be able to capture the dynamics of the thin film. The Bretherton theory stands that thin film is very important as long as is the key point determining the bubble shape. The Figure 2 (left) present the computed shape for a capillary number of 0.01, which is in fully agreement with the theory. The Figure 2 (right) shows the detail of the grid refinement at the rear of the bubble where the film is thinner. We can also note the specific grid refinement along the interface.

![Figure 2](image)

*Figure 2*: (left) Typical bubble in the stationary state. (right) Grid refinement at the rear of the bubble.

The bubble velocity is evaluated by computing the $x$ position of the center mass of the bubble along the channel, \( \frac{\int_S x \ c(x,t) \ dS}{\int_S dS} \) where $x$ is the spatial position of each fraction $c$.

We present now some quantitative and qualitative numerical results on a bubble flowing on a micro channel. For a bubble between parallel plates the analytical solutions are

\[
\frac{U_b}{U_f} \sim 1 + 0.643 \ (3C_a)^{2/3} \quad \text{or} \quad \sim 1 + 0.51 \ (3C_a)^{2/3}
\]

(6)

the second relation valid for very viscous drops (liquid-liquid interfaces). In our simulations we impose a Reynolds number of 0.1 and the only variable parameter is the capillarity number $C_a$. The Figure 3 shows the log-log scaling of the excess of velocity $U_b/U_f - 1$ as function of the capillarity number $C_a$ up to a capillarity number of 0.5 $10^{-4}$. The solid lines shows both limits from the later relations. These results are consistent with those of reference [3] where capillarity number are indeed greater.

Experimental observations in a microfluidic setup [5] show

- that below a capillarity number of around $10^{-5}$ the film gap and the bubble velocity remain constant giving no dependency of the ratios $h/H$ and $U_b/U_f$ on the capillarity number, and
- the bubble lost its Bretherton shape becoming more symmetric, like a pancake.
This phenomenon appears for $h$ of the order of tens of nanometers, an argument advanced as explanation is that in this region, near of the wall, the Van der Waals forces are not negligible pushing away the bubble by the apparition of an equilibrium film of constant width. The main difficult for including the Van der Waals forces in this continuum approach is that’s necessary resolving numerically a lot of scales, from the small, few nanometers for the equilibrium film, to the large ones, the channel width $H$. Including Van der Waals forces in a continuum approach was recently done using Gerris to impose the macroscopic contact angles from microscopic physics \cite{4}, the numerical computations were done locally and the wide scale range problem was not matter of fact. Starting from the Lennard-Jones potential of two particles and doing some approximations we can add to the r.s.h. of the Navier-Stokes momentum equation (equation (2)) the force $F(d)$ per unit of volume which depends only on the distance $d$ between the bubble interface and the wall

$$F(d) = \frac{K}{d^m} \left[ m \left( \frac{h^*}{d} \right)^{m+1} - n \left( \frac{h^*}{d} \right)^{n+1} \right]$$

where $K$ is a constant and $h^*$ is the equilibrium film thickness, $m = 3$ and $n = 2$. (details in reference \cite{4}).

To compare qualitatively this approach without resolving all the spatial scales we impose a large value of $h^*$ which is indeed not physical but the mechanism is still the same, pushing the bubble away from the wall. The Figure 4 present two numerical
simulations of a bubble flowing into a channel for $C_a = 0.01$ and $Re = 0.1$ for the same final time with (upped side) and with out (lower side) Van der Waals forces from equation (7). We observe that (i) imposing the Van der Waals forces we found a larger gap $h$ and consequently a faster bubble velocity $U_b$, (ii) the bubble shape becomes a pancake like.

4 CONCLUSIONS

We have presented numerical simulations of a bubble into a channel, the well behavior of the numerical implementation of the Navier-Stokes equations with a surface tension model was demonstrated by a comparison with Bretherton theory for very small capillary numbers were the scaling law in $C_a^{2/3}$ was validated. The quality of the numerical results are, in particular, a consequence of the grid refinement approach which allows computing the very thin films of liquid between the bubble and the wall. We have also implemented a Van der Waals like force and the imposed minimum gap gives numerical prediction in according with experimental observations.

REFERENCES


GLOBAL REDUCED-ORDER MODEL ADAPTED TO THE LOW- AND MEDIUM-FREQUENCY ANALYSIS OF COMPLEX DYNAMICAL STRUCTURES

O. Ezvan, A. Batou, C. Soize

Université Paris-Est, Laboratoire Modélisation et Simulation Multi-Echelle, MSME UMR 8208 CNRS
5 bd Descartes, 77454 Marne-la-Vallée, France
olivier.ezvan@univ-paris-est.fr

Key words: structural dynamics, reduced-order model, global modes, computational model, frequency range, medium-frequency range

Abstract. In structural dynamics, the use of the vibration eigenmodes (elastic modes) allows for obtaining an accurate small-dimension reduced-order model (ROM) for the low-frequency range analysis. For some complex structures with distinct structural levels (presence of flexible parts attached to a stiff master part), numerous local elastic modes are intertwined with the usual global elastic modes, yielding high-dimension ROM. To circumvent this difficulty, a general method is proposed in order to construct a small-dimension ROM whose reduction vector basis is constituted of the global displacements only. The method is applied to an automobile complex structure.

1 Introduction

In structural dynamics, the low-frequency (LF) range is classically characterized by frequency response functions (FRF) exhibiting well separated resonance peaks that are associated with large-wavelength global shapes. Instead, the high-frequency range, for which the FRF are rather smooth, presents a constant high modal density, which is associated with the presence of numerous elastic modes that are constituted of small-wavelength displacements. An intermediate band, the medium-frequency (MF) band, exists for complex structures [1]. The modal analysis method [2, 3, 4] consists in projecting the dynamics equations onto the elastic modes associated with the first eigenfrequencies. Since in general the LF band presents only a few elastic modes, classical modal analysis is then an efficient tool for constructing a predictive small-dimension reduced-order model (ROM). In this work, we are interested in complex structures for which the modal density can be very high, as soon as low frequencies, due to the presence of numerous local elastic modes intertwined with the usual global elastic modes. This feature is related to the presence of small-scale flexible parts, such as panels, that are attached...
to a stiff master part. Subsequently, such flexible parts are responsible for the presence of elastic modes dominated by local displacements. Furthermore, in such complex structures, small geometrical heterogeneities, although not clearly identified as flexible structural elements, are also responsible for the apparition of numerous local elastic modes.

Recently, some researches [5, 6, 7] have been carried out for dealing with this problematic, whose major objective is to reduce the dimension of the ROM. The present work accounts for the last developments made, which constitute a general framework for the construction of a ROM that is adapted, for the LF and MF bands, to complex structures, and whose reduced dimension is related to its controlled accuracy. Since the contributions of the local displacements are not necessarily significant for predicting the response of the stiff master part (in this work, we are not interested in predicting the small-scale local displacements), we propose a general method for constructing a small-dimension ROM whose reduction vector basis is constituted only of the global displacements. The extraction of the global displacements is based on a filtering strategy that is constituted of two steps. Firstly, an approximation subspace is introduced. This subspace is associated with displacements of reduced kinematics. Secondly, the usual generalized eigenvalue problem, associated with the homogeneous conservative system and spanning the reduction vector basis, is modified: the mass matrix is modified such that it corresponds to the reduced kinematics. It should be noted that, in this unusual eigenvalue problem, the elastic energy is kept exact. The method, whose implementation is well adapted to an efficient use in the context of commercial softwares, is applied to the automobile complex structure introduced in [6].

2 Theory

2.1 Context

We are interested in calculating the frequency response functions (FRF) $U(\omega)$ of a fixed linear structure, for $\omega$ in the frequency band of analysis $B = [\omega_{\min}, \omega_{\max}]$, with $\omega_{\min} > 0$. In the context of finite element analysis [8], vector $U(\omega)$ is the discretization of the displacement field and is the solution, for all $\omega$ in $B$, of the following matrix equation,

$$(-\omega^2 [M] + i \omega [D] + [K]) U(\omega) = F(\omega),$$

where $[M]$, $[D]$, and $[K]$ are the mass, damping, and stiffness matrices, with $F(\omega)$ the discretization of the external forces. These matrices are $(m \times m)$ real positive-definite matrices, where $m$ denotes the number of degrees of freedom (DOF) of the finite element model. The first $n$ elastic modes $\{\varphi^\alpha\}_\alpha$, which are associated with the $n$ smallest eigenvalues $\{\lambda^\alpha\}_\alpha$ such that $\lambda^\alpha = \omega^\alpha_\alpha$, are the solutions of the following generalized eigenvalue problem,

$$[K] \varphi^\alpha = \lambda^\alpha [M] \varphi^\alpha,$$

for obtaining the $(m \times n)$ real matrix $[\Phi] = [\varphi^1 \ldots \varphi^n]$, which constitutes the reduction vector basis used in the classical modal analysis method, for which vector $U(\omega)$ is written, for
in \( B \) and with \( n \ll m \), as

\[
U(\omega) \simeq U^{\text{elas}}(\omega) = \sum_{\alpha=1}^{n} q_{\alpha}(\omega) \varphi^{\alpha} = [\Phi] q(\omega),
\]

in which the \( n \)-dimensional complex vector \( q(\omega) \) of generalized coordinates is the solution of the following reduced-order matrix equation,

\[
(-\omega^2 [M] + i\omega [D] + [K]) q(\omega) = F(\omega),
\]

in which \( F(\omega) = [\Phi]^\top F(\omega) \), and where, for \( A \) in \( \{M, D, K\} \) and \( \Phi \) in \( \{\mathbb{M}, \mathbb{D}, \mathbb{K}\} \), we have \([A] = [\Phi]^\top [\Phi] \).

For the complex structures dealt with, dimension \( n \) for obtaining convergence in \( B \) can be very high. In addition to the increased cost required for computing, in such a case, the first \( n \) elastic modes, the main issue is the high dimension of the classical ROM.

### 2.2 Methodology proposed

The first step of the method consists in introducing, for the kinetic energy, an approximation (reduced kinematics) that is adapted to the filtering of the local displacements. In the next section, the construction of the mass matrix associated with such a reduced kinematics is presented.

#### 2.2.1 Reduced-kinematics mass matrix

Let \( A_g \) denote a given vector subspace of \( \mathbb{R}^m \) and let \( N_g \) denote its dimension. This subspace defines a reduced kinematics. Let also \([B]\) be a \((m \times N_g)\) real matrix whose columns span \( A_g \). For all \( V \) in \( \mathbb{R}^m \), the associated vector \( V^{A_g} \) in \( A_g \) is defined as the orthogonal projection of \( V \) onto \( A_g \). Using the inner-product \( \langle x, y \rangle_M = y^\top [M] x \), the orthogonal-projection matrix \([P_{A_g}]\), which is such that

\[
V^{A_g} = [P_{A_g}] V,
\]

is written as

\[
[P_{A_g}] = [B] \left( [B]^\top [M] [B] \right)^{-1} [B]^\top [M].
\]

Introducing the approximation of Eq. (5) for the kinetic energy, the associated reduced-kinematics mass matrix \([M_{A_g}]\) is written as

\[
[M_{A_g}] = [P_{A_g}]^\top [M] [P_{A_g}].
\]

The \((m \times m)\) real matrix \([M_{A_g}]\) is symmetric positive semidefinite, of rank \( N_g \). This matrix is generally not sparse.
2.2.2 Global-displacements ROM

The first \( n_g \) global eigenvectors \( \{\psi^\alpha\}_{\alpha} \), associated with the \( n_g \) smallest eigenvalues \( \{\sigma_{\alpha}\}_{\alpha} \), are the solutions of the following generalized eigenvalue problem,

\[
[K] \psi^\alpha = \sigma_{\alpha} [M_{\mathcal{A}_g}] \psi^\alpha,
\]

for obtaining the \((m \times n_g)\) real matrix \([\Psi] = [\psi^1 \ldots \psi^{n_g}]\), which constitutes the reduction vector basis used in the proposed method, for which vector \( \mathbb{U}(\omega) \) is written, for \( \omega \in \mathcal{B} \) and with \( n_g \leq n \), as

\[
\mathbb{U}(\omega) \simeq \mathbb{U}_{\text{glob}}(\omega) = \sum_{\alpha=1}^{n_g} q^\alpha_{\omega}(\omega) \psi^\alpha = [\Psi] \mathbf{q}^\omega(\omega),
\]

and in which the \( n_g \)-dimensional complex vector \( \mathbf{q}^\omega(\omega) \) of generalized coordinates is the solution of the following small-dimension matrix equation,

\[
(-\omega^2 [M^g] + i\omega [D^g] + [K^g]) \mathbf{q}^\omega(\omega) = \mathcal{F}^g(\omega),
\]

in which \( \mathcal{F}^g(\omega) = [\Psi]^T \mathbb{F}(\omega) \), and where, for \( A \) in \{\( M, D, K \)\} and \( \mathcal{A} \) in \{\( \mathbb{M}, \mathbb{D}, \mathbb{K} \)\}, we have \([A^g] = [\Psi]^T [\mathcal{A}] [\Psi]\).

2.2.3 Computational aspects

Although \((m \times m)\) matrix \([M_{\mathcal{A}_g}]\) is full, its assembly can be avoided by using a subspace iteration algorithm for solving Eq. (8). Nevertheless, access to stiffness matrix \([K]\) may not be possible in the context of the use of commercial softwares. In order to circumvent both these difficulties, a method (that we will call the indirect method) is proposed. For all \( \alpha \) in \{1, \ldots, \( n_g \)\}, eigenvector \( \psi^\alpha \) is approximated such that

\[
\psi^\alpha = [\Phi] s^\alpha,
\]

in which \( \{s^\alpha\}_{\alpha} \) are \( n_g \)-dimensional real vectors, which are the solutions of the following reduced-order generalized eigenvalue problem,

\[
[A] s^\alpha = \sigma_{\alpha}[M_{\mathcal{A}_g}] s^\alpha,
\]

in which \([A] = [\Phi]^T [K] [\Phi] \) and \([M_{\mathcal{A}_g}] = [\Phi]^T [M_{\mathcal{A}_g}] [\Phi] \). Since the contributions of the discarded local displacements are neglected, a residual error is introduced in the global-displacements ROM. A convergence analysis with respect to subspace \( \mathcal{A}_g \) allows for reaching a compromise between dimension \( n_g \) and accuracy of the ROM. Using the indirect method, the overall numerical cost for performing such a convergence analysis is then nearly reduced to the computation of the \( n \) elastic modes.
3 Application to an automobile complex structure

We use the indirect method proposed, in order to construct a small-dimension ROM for the automobile complex structure whose computational model is displayed in Fig. 1 (in which the gray intensity is related to the level of rigidity).

![Computational model](image1.png)

**Figure 1:** Computational model, with location of excitation nodes exc1 and exc2, and of observation node obs. Gray intensity is related to the level of rigidity (the darker is the stiffer).

There are well identified flexible parts (such as the roof and the floor panels) as well as numerous structural heterogeneities (see for example in the front of the car) that are both responsible for the presence of numerous local elastic modes. The computational model has \( m = 1,462,698 \) DOF and there are 1,048 elastic modes in the frequency band of analysis \( B = 2\pi \times [0, 500] \) rad/s, while the classical ROM is converged with \( n = 1,457 \). Figure 2 displays, on the left, a purely local elastic mode and, on the right, a purely global elastic mode.

![Elastic modes](image2.png)

**Figure 2:** Left: elastic mode \( \varphi^1 \) (24 Hz), right: elastic mode \( \varphi^3 \) (39 Hz). Gray intensity is related to the level of amplitude of the displacements (the greater amplitude is the lighter)

In the method proposed, the global-displacements basis, represented by matrix \([ \Psi ]\), is entirely
defined upon the subspace $\mathcal{A}_g$ chosen, as well as upon truncation order $n_g$. In this application, $\mathcal{A}_g$ is spanned by vectors that consist in discrete multivariate Legendre polynomials whose degree, $N_d$, allows for controlling the filtering of the local displacements. The ROM convergence is analyzed in studying the FRF obtained under the application of unit forces in the $x$- and $z$-directions and unit moments around the $x$- and $y$-axes, relative to the excitation nodes exc1 and exc2 (depicted in Fig. 1). This convergence analysis of the global-displacements ROM is done with respect to $N_d$, and, for fixed $N_d$, with $n_g$ chosen as the smallest value such that the associated highest eigenfrequency is greater than 525 Hz. Figure 3 displays the FRF of observation node obs (norm of the displacements of the node, in log scale) obtained using Eq. (2) for the reference, and using Eq. (9) for the small-dimension ROM in which $N_d$ is successively chosen as equal to 5, 10, 15, and 20. The resulting dimensions $n_g$ are 168, 355, 479, and 624.

![Figure 3: Logarithm of the norm of the displacements (m) versus frequency (Hz): reference $\|u_{\text{obs}}^\text{elas}\|$ (black solid line), ROM response $\|u_{\text{obs}}^\text{glob}\|$ for $N_d$ equal to either 5, 10, 15, or 20 (gray dashed line).]

It can be seen in Fig. 3 that the differences with respect to the reference are decreasing with respect to the increase of $N_d$ (and thus $n_g$). The differences are larger in the high part of the frequency band, while it is known that the experimental variabilities are generally larger in this band, too. For $n_g = 479$ or $n_g = 624$, the error introduced in the computation of the response is not as important as the experimental variabilities may be [9, 10].

4 Conclusions

A general and efficient method for constructing a reduced-order computational model, which is adapted to predicting the global displacements of complex structures for which there are numerous local elastic modes intertwined with the global elastic modes, has been presented.
The strategy relies on the filtering of the local displacements in order to construct a small-dimension ROM. The filtering is obtained by reducing the kinematics of the kinetic energy in the usual eigenvalue problem. A convergence analysis with respect to the reduced-kinematics responsible for the filtering of the local displacements allows to control, with a reduced cost, the compromise between the ROM accuracy and dimension.

REFERENCES


ESSENTIAL BOUNDARY CONDITIONS AND K-COMPRESSIBILITY IN HYPER REDUCED ORDER MODELS

ALEJANDRO COSIMO\textsuperscript{1}, ALBERTO CARDONA\textsuperscript{1} AND SERGIO IDELSOHN\textsuperscript{1,2}

\textsuperscript{1}Centro de Investigación de Métodos Computacionales (CIMEC)
Santa Fe, Argentina

\textsuperscript{2}International Center for Numerical Methods in Engineering (CIMNE) and Institució Catalana de Recerca i Estudis Avançats (ICREA)
Barcelona, Spain

Key words: HROM, FEM, Essential Boundary Conditions

Abstract. The treatment of inhomogeneous essential boundary conditions must be carefully analysed in the formulation of Hyper Reduced Order Models. Two methods are analysed: one in which the boundary conditions are imposed in an strong way, and a second one in which a weak imposition of boundary conditions is made. Applications to nonlinear transient heat conduction problems with temperature dependent thermophysical properties and time dependent essential boundary conditions are considered.

1 INTRODUCTION

Problems whose complexity is intractable from the point of view of real time applications are quite frequently found in science and in engineering. Nowadays, the design and formulation of Reduced Order Models for nonlinear problems is becoming popular in the computational mechanics community, as an elegant and promising solution to this problem.

Let $\mathcal{S}^h \subset \mathcal{S}$ and $\mathcal{V}^h \subset \mathcal{V}$ be subspaces of the trial $\mathcal{S}$ and test $\mathcal{V}$ functional spaces used in the definition of a variational problem. Generally, in the formulation of Reduced Order Models (ROMs), an approximate solution $\hat{T}^h$ to $T^h \in \mathcal{S}^h$ is sought in a subspace of $\mathcal{S}^h$ by defining a new basis $\hat{X} \in \mathbb{R}^{N \times k}$, where $N$ is the number of degrees of freedom (DOFs) of the High Fidelity (HF) model and $k$ is the dimension of the basis spanning the subspace of $\mathcal{S}^h$. If a Bubnov-Galerkin projection is used, approximate versions $\hat{w}^h \in \text{span}\{X\}$ of the test functions $w^h$ are built, and functions $\hat{T}^h \in \mathcal{S}^h$ are approximated by affine translations of the test functions $\hat{w}^h$. In POD based ROMs, the new basis $X$ is built by
computing the Singular Value Decomposition (SVD) of a set of snapshots that are given by time instances of the spatial distribution of the solution to a training problem. It is known that the vectors comprising this basis inherit the behaviour of the elements of the snapshots’ set, hindering the possibility of reproducing non-admissible test functions. That is why, careful attention must be paid on how the snapshots for building $X$ are collected. This issue is studied in detail in this work.

The concept of consistent snapshots collection procedures for nonlinear problems was first introduced by Carlberg et al. [1, 2]. As they pointed out in [2] “most nonlinear model reduction techniques reported in the literature employ a POD basis computed using as snapshots $\{T_n|n = 1, \ldots, n_t\}^1$, which do not lead to a consistent projection”. In the last expression $n_t$ is given by the number of time steps comprising the training problem and $T_n$ are the parameters such that $T^n = NT_n$ with $N$ given by the shape functions used for interpolation. The lack of consistency of formulations that use the collection procedure described before, finds its roots in the fact that when computing the POD basis with time instances of $T^h$, that is by $\{T_n|n = 1, \ldots, n_t\}$, if non-zero essential boundary conditions are present, $\text{span}\{X\} \not\subset V^h$ because any $v \in \text{span}\{X\}$ is not identically zero at the portion of the boundary with non-homogeneous essential conditions.

Carlberg et al. [2] proposed two alternative procedures to collect snapshots for $T^h$:

- Snapshots of the form $\{T_n - T_{n-1}|n = 1, \ldots, n_t\}$. The problem with this strategy is that the set of snapshots is characterised by high frequency content, giving a less compressible SVD spectrum [3] than when using a collection procedure based on the snapshots of the solution. Another problem of this strategy are the essential boundary conditions that depend on time. In this case, it cannot be guaranteed that the snapshots will be identically zero at the boundary with essential boundary conditions.

- Snapshots of the form $\{T_n - T_0|n = 1, \ldots, n_t\}$, where $T_0$ is the initial condition. With this strategy it cannot be guaranteed that functions in $\text{span}\{X\}$ will be test functions, for instance, when essential boundary conditions are different from $T^h_0$. Amsallem et al. [4] have observed that this strategy leads to more accurate ROMs than the previous strategy. As it is discussed in that work, using a different initial condition $T^*_0$ in the online stage requires in principle recomputing the snapshots for reconstructing the POD modes for projection. Several fast alternatives to solve this problem are proposed in [4].

Pinnau [5] discussed the alternative of using a snapshots collection of the form $\{T_n - T_m|n = 1, \ldots, n_t\}$, where $T_m$ is given by the mean value of $\{T_n|n = 1, \ldots, n_t\}$. Likewise to Carlberg’s, this strategy suffers of the problems mentioned before.

---

1 For the sake of conciseness, in this work we do not consider the objective function $T^h$ to depend on a set of analysis parameters $\mu$. If this were the case, the snapshots collection strategies introduced herein apply directly just by applying them to each of the training parameters $\mu_i$. 

1030
Gunzburger et al. [6] presented two schemes for handling inhomogeneous essential boundary conditions in the context of ROMs. They assumed that the Dirichlet boundary is divided into a set of $P$ non-overlapping portions where the involved field is imposed as $\beta_p(t)g_p(x)$, for $p = 1, \cdots, P$, where $g_p$ are given functions and $\beta_p$ are time dependent parameters. In a first approach, the solution is written in terms of a linear combination of test functions vanishing in the portion of the boundary with essential boundary conditions, and in terms of a linear combination of particular solutions of the steady state version of the problem to be solved. In a second approach, they proposed to express the solution in terms of a set of POD basis functions not vanishing on the Dirichlet boundary, and adding a set of equations describing the essential boundary condition. Then, they use the QR decomposition on the resulting system in order to obtain a set of test functions vanishing on that boundary. These techniques proved to work well in the context of ROMs [6].

In this work, we propose and test different alternatives for dealing with general inhomogeneous essential boundary conditions, including the possibility of time dependence. Although the strategies introduced in this work are of widely general application, the presented ideas are applied to nonlinear transient heat conduction problems with thermophysical properties depending on temperature. In section 2, we present the problem statement and two variational formulations, one weakly imposing Dirichlet boundary conditions and the other strongly imposing those conditions. In section 3, an HROM that considers strong enforcement of boundary conditions is presented. In section 4, the formulation of two HROMs that weakly impose Dirichlet conditions is introduced. Several examples are presented in section 5 to assess the performance of the proposed solutions. Finally, conclusions are given.

2 PROBLEM STATEMENT, VARIATIONAL FORMULATION AND FINITE ELEMENT DISCRETISATION

The physical problem under consideration is a nonlinear transient heat conduction problem, with temperature dependent thermophysical properties. The problem is described by the equation

$$\rho c \dot{T} = Q + \nabla \cdot (k \nabla T) \quad \forall (x, t) \in \Omega \times (t_0, \infty)$$  \hspace{1cm} (1)

where $\rho$ is the density, $k$ is the thermal conductivity, $c$ the heat capacity, $T$ is the temperature, $Q$ is the external heat source per unit volume, and $\Omega$ is the space domain. The temperature field should verify the initial condition $T(x, t = 0) = T_0$ \(\forall x \in \Omega\), where $T_0$ is the given initial temperature field. Additionally, the following set of conditions must be verified at the disjoint portions $\Gamma_d, \Gamma_q, \Gamma_c$ of the external boundary: $T\big|_{\Gamma_d} = T_d$, $k \nabla T \cdot n\big|_{\Gamma_q} = q_w$, and $k \nabla T \cdot n\big|_{\Gamma_c} = h_f(T_f - T)$, where $\Gamma_d \cup \Gamma_q \cup \Gamma_c = \partial \Omega$, and where $T_d$ is the imposed temperature at the boundary $\Gamma_d$, $q_w$ is the external heat flow at the boundary $\Gamma_q$, $h_f$ is the heat convection coefficient, $T_f$ is the external fluid temperature at the portion the boundary $\Gamma_c$ and $n$ is the outward normal to the boundary under consideration.
In what follows, we briefly present the variational formulation of the problem and its finite element discretisation. Essential boundary conditions can be enforced strongly or weakly. In order to strongly enforce Dirichlet boundary conditions, let $S = \{ T \in H^1(\Omega) / T|_{\Gamma_e} = T_d \}$ be the space of trial solutions and $V = \{ v \in H^1(\Omega) / v|_{\Gamma_d} = 0 \}$ be the space of weighting or test functions, where $H^1$ is the first order Sobolev space. Then, the variational formulation is given as follows: Find $T \in S$ such that $\forall w \in V$

$$\int_{\Omega} w \left[ \rho c \frac{\partial T}{\partial t} - Q \right] d\Omega + \int_{\Omega} \nabla w \cdot (k \nabla T) d\Omega + \int_{\Gamma_c} wh_f(T - T_f) d\Gamma$$

$$- \int_{\Gamma_q} w q_w d\Gamma = 0, \quad \text{for } t > 0;$$

$$\int_{\Omega} w T d\Omega = \int_{\Omega} w T_0 d\Omega, \quad \text{for } t = 0. \quad (2)$$

Let $S^h \subset S$ and $V^h \subset V$ be subspaces of the trial and test functional spaces. Then, in matrix notation, $T^h \in S^h$ is given by $T^h(\mathbf{x}, t_n) = \mathbf{N}^T \mathbf{T}_n$, where $\mathbf{N}$ denotes the finite element basis and $\mathbf{T}_n \in \mathbb{R}^N$ denotes the Finite Element Method (FEM) degrees of freedom, with $N$ the dimension of the FEM space. Then, using a Bubnov-Galerkin projection and a modified Backward-Euler scheme for time integration, the residual of the nonlinear thermal problem in its discrete expression reads [3]

$$\Pi_n = H_n^c - H_n^{c-1} \Delta t + G_n^k + F_n - Q_n = 0, \quad (3)$$

where

$$G_n^k = \left( \int_{\Omega} \nabla N_k \nabla N^T \Omega + \int_{\Gamma_c} h_f N N^T d\Gamma \right) \mathbf{T}_n, \quad (4)$$

$$F_n = - \int_{\Gamma_q} N_q w_n d\Gamma - \int_{\Gamma_c} h_f N T_{f_n} d\Gamma, \quad (5)$$

$$Q_n = \int_{\Omega} \mathbf{N} Q_n d\Omega, \quad (6)$$

$$H_n^c = \int_{\Omega} \rho c_n \mathbf{N} \mathbf{N}^T d\Omega \mathbf{T}_n. \quad (7)$$

In order to weakly impose Dirichlet boundary conditions, the use of Lagrange multipliers is adopted. The idea is to remove from the trial and test function spaces, the constraint over the portion of the boundary corresponding to essential boundary conditions. Accordingly, let $V = \{ v \in H^1(\Omega) \}$ be the space of trial and test functions for the temperature, and let $Q = \{ q \in L_2(\Gamma) \}$ be the space of trial and test functions for the Lagrange multipliers. Then, the variational formulation is given as follows: Find
$$(T, \lambda) \in V \times Q$$ such that $\forall (w, q) \in V \times Q$

\[
\int_\Omega w \left[ \frac{\partial T}{\partial t} - Q \right] d\Omega + \int_\Omega \nabla w \cdot (k \nabla T) d\Omega + \int_{\Gamma_c} w h_f (T - T_f) d\Gamma - \int_{\Gamma_q} w q d\Gamma + \int_{\Gamma_d} w \lambda d\Gamma + \int_{\Gamma_d} q (T - T_d) d\Gamma = 0, \quad \text{for } t > 0;
\]

\[
\int_\Omega w T d\Omega = \int_\Omega w T_0 d\Omega, \quad \text{for } t = 0.
\] (8)

As similarly done before, let $V_h \subset V$ and $Q_h \subset Q$. Therefore, in matrix notation, $T^h \in V_h$ and $\lambda^h \in Q_h$ are given by $T^h(x, t_n) = N^T T_n$ and $\lambda^h(x, t_n) = \bar{N}^T \lambda_n$ where $N$ denotes the finite element basis for the temperature field, and $T_n \in \mathbb{R}^N$ denotes the FEM nodal degrees of freedom. Similarly, $N$ denotes the finite element basis for the Lagrange multipliers, and $\lambda_n \in \mathbb{R}^{N\lambda}$ denotes the parameters corresponding to the Lagrange multipliers. Then, the residual characterising the FEM discretisation can be written as

\[
\Pi_n = \begin{bmatrix} \Pi_{T_n} \\ \Pi_{\lambda_n} \end{bmatrix} = \begin{bmatrix} H_n - H^c_n - \frac{1}{\Delta t} + C^k_n + F_n + B^\lambda_n - Q_n \end{bmatrix} = 0,
\] (10)

where the new terms with respect to the previous formulation are given by

\[
B^\lambda_n = B^\lambda_n(\lambda_n, t_n) = \int_{\Gamma_d} \bar{N} \bar{N}^T d\Gamma \lambda_n
\] (11)

\[
B^{T_d}_{n} = B^{T_d}_{n}(T_n, t_n) = \int_{\Gamma_d} \bar{N} \bar{N}^T d\Gamma T_n - \int_{\Gamma_d} \bar{N} T_d d\Gamma.
\] (12)

3 HROM FORMULATION BY STRONGLY ENFORCING BOUNDARY CONDITIONS

The HROM associated to the formulation given by equation (3) is here introduced. Each non-linear contribution to $\Pi_n$ is hyper-reduced separately as done by Cosimo et al. [3]. Therefore, each of these terms has associated a particular POD basis for its gappy data reconstruction. In what follows, suffices $i \in \{c, k, f, q\}$ are used to identify the different POD bases $\Phi_i$. We emphasise that the sampling is performed independently for each term, but the number of sampling points $n_s$ and the number of gappy modes $n_g$ are always the same for all of them. In what follows, $\tilde{}$ denotes the vector of $n_s$ components sampled from the associated complete term. To compute the POD modes $\Phi_i$, snapshots are taken for each individual contribution at each time step, after convergence of the Newton-Raphson scheme.

To obtain the hyper-reduced residual $\Pi^p_n$ we project the gappy approximation to $\Pi_n$ with the basis $X$, obtained as shown below. From this we have

\[
\Pi^p_n = A_c \tilde{H}^c_n - \frac{\tilde{H}^c_n - \tilde{H}^c_{n-1}}{\Delta t} + A_k \tilde{G}^k_n + A_f \tilde{F}_n - A_q \tilde{Q}_n,
\] (13)
where $A_i = X^T \Phi_i (\tilde{\Phi}_i^T \tilde{\Phi}_i)^{-1} \tilde{\Phi}_i^T$, with $i \in \{c, k, f, q\}$. Note that matrices $A_i$ are computed in the off-line stage.

In what follows, a consistent snapshot collection strategy for $X$ taking into account general essential boundary conditions is introduced. When solving the variational problem given by equation (2) in a finite dimensional space, an approximate solution $T^h \in \mathcal{S}^h$ is described as $T^h = T^h_d + v^h$, where $v^h \in \mathcal{V}^h$ and $T^h_d$ is the finite dimensional version of $T_d$. Then, the trial solutions $T^h$ and the test functions $w^h$ are given by

$$
T^h = T^h_d + v^h = N^I^T T^I_n + N^B^T T^B_n, \tag{14}
$$

$$
w^h = N^I^T \eta, \tag{15}
$$

where $\eta$ are the parameters associated to the test functions and the DOFs $T_n$ are discriminated in terms of parameters describing the boundary with Dirichlet boundary conditions, $T^B$, and the DOFs $T^I$ that are not part of that boundary. Functions $N^I$ and $N^B$ are the FEM shape functions associated to the internal and boundary DOFs, respectively.

Functions with global support are used in the context of ROMs, in contrast to FEM basis functions whose support is local. Therefore, the notion of internal/boundary degrees of freedoms is lost in ROMs, making it necessary to express $T^h$ and $w^h$ as

$$
T^h \simeq \tilde{T}^h = T^h_d + N^T X a_n, \tag{16}
$$

$$
w^h \simeq \tilde{w}^h = N^T X w_n, \tag{17}
$$

where $a_n$ and $w_n$ are the amplitudes associated to the modes $X$.

In order to get admissible test functions $\tilde{w}^h$, the restriction $\tilde{w}^h|_{d} = 0$ must be satisfied. That is why, for the design of a consistent snapshot collection strategy, the snapshots must be of the form $T - T_d$. Then, the problem resides in the correct description of $T^h_d$. A naive solution would be to describe it as in standard FEM, i.e., $T^h_d = N^B^T T^B_n$, but this would lead to a snapshots set with a very high frequency content, decreasing the compressibility of the signal [3].

In order to avoid this inconvenience, we propose to compute a set of static modes that describe the behaviour of the portion of the boundary with essential boundary conditions. The procedure is similar to that followed by the Craig-Bampton or by the Guyan-Irons methods [7]. As we want to build a set of static modes to describe the boundary, we consider only the term $G^k$ in equation (3). Simplifying notation, this term at time instant $t_n$ is given by $G^k = K T$, where $K$ is any linearisation of the stiffness matrix and $T \equiv T_n$. We can neglect non-linearities at this point because we are only interested in finding a basis for expressing the essential boundary conditions. Then, by partitioning in internal and boundary DOFs, we write:

$$
G^k = K T = \begin{bmatrix} K_{II} & K_{IB} \\ K_{BI} & K_{BB} \end{bmatrix} \begin{bmatrix} T^I \\ T^B \end{bmatrix} = 0, \tag{18}
$$
from which we get by static condensation $T^I = -K^{-1}_{II}K_{IB}T^B$. In this case $T^I$ can be regarded as the response to an imposed temperature $T^B$ in the portion of the boundary where only the term $G^k$ is considered. Then, the static modes, which describe the response to unit temperatures imposed at $\Gamma_d$, are given by

$$\Phi_B = \begin{bmatrix} -K^{-1}_{II}K_{IB} \end{bmatrix},$$

and the function $T^h_{\Gamma_d}$ used to denote the essential boundary conditions is assumed to lie in $\text{span}\{\Phi_B\}$. We remark that this procedure is similar to the method proposed by Gunzburger et al. [6] that considers particular solutions derived from the steady state version of the system of equations, but interpreted from a different perspective.

Then, the approximation $\hat{T}^h$ is given by $\hat{T}^h = NT\Phi_BT_n^B + NTXa_n \simeq T^h$. Note that the static modes $\Phi_B$ have the property to be interpolatory at the boundary $\Gamma_d$, thus $T_n^B$ has the physical interpretation to be the value of the field at the nodes lying on $\Gamma_d$.

From this equation the following snapshots collection procedure arises: once the static modes were computed, take snapshots of the form $S_p = \{T_n - \Phi_BT_n^B|n = 1, \cdots, n_t\}$. This strategy has these advantages

- The snapshots given by $S_p$ tend to preserve the $k$-compressibility posed by the field $T^h$.
- General essential boundary conditions can be represented by $\Phi_B$, while keeping simple the process of imposing essential boundary conditions because of the interpolatory property of $\Phi_B$ at $\Gamma_d$.
- Using different initial conditions in the online stage does not require recomputing the snapshots for $X$ or considering another alternative.

4 HROM FORMULATION BY WEAKLY ENFORCING BOUNDARY CONDITIONS

Two alternative HROMs associated to the formulation given by equation (10) are now introduced, aimed at reducing the temperature DOFs $T_n$ and Lagrange multipliers $\lambda_n$. In the first one, $[T_n; \lambda_n]$ is reduced as a unit like $[T_n; \lambda_n] = X^ec_n$, where the POD modes $X^e$ are built from a set of snapshots composed by the temperature field and the Lagrange multipliers, and $c$ denote the associated parameters. A second alternative is to reduce each physical quantity separately like $T_n = Xa_n$ and $\lambda_n = Yb_n$, where $X$ and $a_n$ are the POD modes and the parameters associated to the temperature field, and $Y$ and $b_n$ are the POD modes and the parameters associated to the Lagrange multipliers. From a general point of view, weakly enforcing boundary conditions would have the following advantages with respect to the use of static modes to represent essential boundary conditions:

- Test functions for the temperature field are not required to meet the constraint $T|\Gamma_d = 0$. 
• For problems requiring a large number of static modes, the cost of computing the product $\Phi B T_n^B$ can be a penalising factor. By using Lagrange multipliers, this problem is avoided.

When adopting the first option, the residual $\Pi_n$ given by equation (10) is projected to the space spanned by $X^c$ and each term is separately hyper-reduced as done in [3], and the expression is quite similar to the one given by equation (13) but taking into account the terms involving the restriction over the Dirichlet boundary.

In the second approach proposed in this section, each term of the residual $\Pi_n$ from equation (10) is projected separately according to

$$
\begin{bmatrix}
\Pi_{T_n}^p \\
\Pi_{\lambda_n}^p
\end{bmatrix} =
\begin{bmatrix}
X^T \Pi_{T_n} \\
Y^T \Pi_{\lambda_n}
\end{bmatrix}
$$

(20)

Then, again, each contribution is separately hyper-reduced following the work of Cosimo et al. [3].

5 APPLICATION EXAMPLES

We will show the application of the proposed snapshot collection strategies to two non-linear transient heat conduction problems, with time dependent essential boundary conditions. To assess the performance and robustness of the proposed methods, we study the relative error introduced by the HROM. The relative error $\epsilon$ characterising the HROM as a function of time is measured as

$$
\epsilon = \frac{\|T_R - T_H\|_{\max}}{\|T_H\|}
$$

where $T_R$ is the solution obtained with the HROM, $T_H$ is the High Fidelity solution for same problem and $\| \cdot \|$ denotes the $L_2$ norm. Tri-linear hexahedral elements are used in the examples to interpolate the temperature field. The Lagrange multipliers field is interpolated with bi-linear quadrilateral elements. In what follows, $n_p$ is used to denote the number of POD modes for $T_n$, and $n_\lambda$ is used to denote the number of POD modes for $\lambda_n$.

5.1 Example 1

This example has been presented by Gunzburger et al. [6]. It consists in a linear transient heat conduction problem with $\rho = k = c = 1$ and with a non-linear heat source equal to $Q(T) = -T^2$. The domain to be analysed is a $1 \times 1 \times 0.1428$ cuboid. It is discretised using tri-linear hexahedral elements with a total of 675 degrees of freedom. A time step $\Delta t = 0.01$ is used for the time interval $[0, 1]$. The body is initially at temperature $T_0 = 0$. A time dependent essential boundary condition, $T_d(x, t)$, is imposed, given by

$$
T_d(x, t) \equiv T_d(x, y, z, t) =
\begin{cases}
2t4x(1 - x) & \text{if } y = 1 \land 0 \leq t < 0.5, \\
2(1 - t)4x(1 - x) & \text{if } y = 1 \land 0.5 \leq t \leq 1, \\
4(t - t^2)4x(1 - x) & \text{if } y = 0 \land 0 \leq t \leq 1, \\
|\sin(2\pi t)|4y(1 - y) & \text{if } x = 0 \land 0 \leq t \leq 1, \\
|\sin(4\pi t)|4y(1 - y) & \text{if } x = 1 \land 0 \leq t \leq 1,
\end{cases}
$$

1036
The sides $z = 0$ and $z = 0.1428$ of the domain are insulated. A total of 20 gappy points and 20 gappy modes were used in all cases.

Figure 1: Relative errors for the temperature and Lagrange multipliers fields.

The error obtained using static modes to represent the essential boundary condition can be observed in Figure 1a, where different numbers of projection modes were considered. As it can be seen, good results are obtained.

We have two alternatives in the case of weakly imposing the essential boundary conditions. The error behaves as shown in Figure 1b when reducing $T_n$ and $\lambda_n$ as a unit. Although the obtained results seem to be quite good, we remark that convergence is not achieved for the cases $n_p < 7$, $np = 10$ and $n_p = 11$. Monotone convergence, for any number of projection modes, is achieved only when using 12 or more modes. We suppose
that this behaviour is related to the fact that the temperature field must have enough freedom to be able to meet the restrictions imposed by the Lagrange multipliers.

The error obtained when reducing \( T_n \) and \( \lambda_n \) separately can be observed in Figure 1c. In these tests, we took \( n_\lambda = 4 \). It should be kept in mind that \( n_p \) should be greater than \( n_\lambda \), otherwise \( T_n \) will not have enough freedom to satisfy the restrictions. In this case, convergence can be achieved for \( n_p \geq 4 \), but a good approximation error to the temperature field is observed for \( n_p \geq 7 \). We remark that when \( n_\lambda > 4 \) in this numerical experiment, a bad conditioning of the reduced iteration matrix was obtained. A pivoting strategy was used to get convergence, with elimination of the equations associated to zero pivots, and it was observed that most of the time the constraint equations corresponding to modes higher than four were eliminated.

When comparing the three different alternatives, it is observed that the lowest error option is when using static modes. Nevertheless, the cost is higher than in the strategies that impose the Dirichlet boundary conditions weakly. Concerning the two latter alternatives, it is observed that the strategy of reducing \( T_n \) and \( \lambda_n \) as a unit leads to the lowest errors of both for the temperature, for the same number of reduced DOFs. For example, when using that alternative with \( n_p = 12 \) the error is \( O(10^{-4}) \), but when reducing \( T_n \) and \( \lambda_n \) separately with \( n_p = 8 \) and \( n_\lambda = 4 \), the error is \( O(10^{-3}) \). The approximation error to \( \lambda_n \) is always lower when reducing \( T_n \) and \( \lambda_n \) as a unit. However, this method is less robust than separate reduction, because in the latter case convergence is achieved for a few number of POD modes of the temperature field whereas the former method was not able to converge in this case.

5.2 Example 2

We consider next a non-linear transient heat conduction problem, where the heat capacity is \( c = 0.1792 \ T + 495.20 \) and the thermal conductivity is \( k = 0.25 \ T + 70 \). The material density is \( \rho = 1 \). The domain to be analysed is a \( \pi \times \pi \times 0.4487 \) cuboid. It is discretized using tri-linear hexahedral elements with a total of 675 degrees of freedom. A time step \( \Delta t = 1 \) is used for the time interval \([0, 600]\). The body is initially at temperature \( T_0 = 1200 \). A time dependent essential boundary condition is imposed on side \( x = 0 \). The other sides of the domain are insulated. The essential boundary condition \( T_d(x, t) \) is given by

\[
T_d(x, t) = T_0 \frac{e^{-t/600} - 1}{e^{-1} - 1} \cos \left( \frac{\gamma_t}{2} + \frac{\pi t}{300} \right) \cos(\pi \gamma) + T_0, \quad \gamma_t = \frac{26\pi t}{600} + \frac{y}{2}.
\]

A total of 60 gappy points and 60 gappy modes were used in all cases, except for the equations corresponding to the Lagrange multipliers when reducing \( T_n \) and \( \lambda_n \) separately, where 5 gappy points and modes were used.

The results obtained for the different schemes can be observed in Figures 2a, 2b and 2c. Similar comments as in the previous example apply in this case. We remark that the
scheme that weakly imposes the Dirichlet boundary conditions and that reduces $T_n$ and $\lambda_n$ as unit, begins to converge for $n_p \geq 13$.

6 CONCLUSIONS

Several alternatives were analysed for building Hyper-Reduced Order Models to solve nonlinear thermal problems with time dependent inhomogeneous essential boundary conditions. One strategy considers the use of static modes for strongly imposing Dirichlet boundary conditions, proving to be quite efficient in terms of the obtained error. However, when the number of static modes increased, the computational cost also increased. The other considered alternatives imposed weakly the Dirichlet boundary conditions. It was observed that the strategy of reducing $T_n$ and $\lambda_n$ as a unit lead to the lowest errors of
both for the temperature, for the same number of reduced DOFs. However, this method was less robust than separately reducing the temperature and Lagrange multipliers field.

REFERENCES


NUMERICAL SIMULATION OF ULTRASONIC WAVE PROPAGATION USING HIGHER ORDER METHODS IN SPACE AND TIME

Jörg F. Unger

1 Federal Institute for Materials Research and Testing, Unter den Eichen 87, 12205 Berlin, Germany, joerg.unger@bam.de

Key words: elastic wave propagation, spectral element method, Nyström methods.

Abstract. The paper discusses the efficient simulation of ultrasonic wave propagation. It is demonstrated that a combination of higher methods in space and time leads to a significant performance boost. Higher order spectral elements are used for the spatial discretization. A comparison with standard finite elements shows the advantages when using explicit time integration schemes. For the temporal discretization, an efficient explicit fourth order Nyström method is presented. Its computational efficiency for wave propagation problems is compared to a second order Velocity Verlet integration.

1 Motivation

The applications dealing with the numerical simulation of elastic waves range from large scale seismic problems [1] to small scale ultrasonic testing methods, which belong to the group of nondestructive testing methods. They are widely used to detect damage or defects in engineering materials such as concrete and are based on the characteristics of propagating elastic waves. All of the methods are usually based on determining the traveling time of an elastic wave from its origin (external pulse or the formation of a crack) to a sensor. Assuming some degree of heterogeneity of the system, the average wave speed can be determined, which is related to the local stiffness. While monitoring the specimen during its lifetime, changes of the propagation of ultrasonic waves can then be related to the evolution of damage. Unfortunately, usual loading scenarios are complex and do not lead to a homogeneous distribution of damage. In addition, the specimens under investigation can have complex geometries. As a consequence, it is often required to simulate the wave propagation numerically in order to interpret the experimental data.

The focus of the paper is the efficient numerical simulation of ultrasonic wave propagation using higher order methods in space and time. It is shown that only a combination of higher order methods in the spatial as well as in the temporal domain yields efficient solution schemes. After a short introduction, the spatial discretization is discussed with
a short review on the spectral element method. Afterwards, a temporal discretization using a second order Velocity Verlet and a fourth order Nyström method is proposed. In the last section, the accuracy and computational efficiency of the spatial and temporal discretization is analyzed.

2 Introduction and problem formulation

Using the equilibrium equations, the assumptions of small strains and a linear elastic material behavior, which is motivated by the small displacement/stress amplitudes due to the external pulse, the governing equations are given as

\[ 0 = -\rho \ddot{u} + \nabla \cdot \sigma \quad \varepsilon = \frac{1}{2} (\nabla u + \nabla u^T) \quad \sigma = C\varepsilon, \quad (1) \]

with the density \( \rho \), the displacements \( u \), the strains \( \varepsilon \), the stresses \( \sigma \) and the elasticity tensor \( C \). The governing equations are complemented by initial conditions and boundary conditions

\[ u(x,0) = u_0 \quad x \in \Omega, \quad \sigma n = t(t) \quad x \in \Gamma^f, \quad \dot{u}(x,0) = v_0 \quad x \in \Omega, \quad u(x,t) = \bar{u}(t) \quad x \in \Gamma^u, \quad (2) \]

with initial displacements and velocities \( u_0 \) and \( v_0 \), the external stress vector \( t(t) \) and the prescribed displacements \( \bar{u}(t) \). The numerical solution is obtained by discretizing the problem in space and time.

The spatial discretization of the governing equations results in a semi-discrete wave equation of the form

\[ M\ddot{d} + Kd = F^e(t), \quad (3) \]

with the mass matrix \( M \), the stiffness matrix \( K \), the external force \( F^e(t) \) and the nodal displacements \( d \), which has to be discretized in time.

3 Spatial discretization and the spectral element method

In the absence of analytical solutions to the wave equations, numerical methods are required to obtain an approximate solution. This problem often occurs in geophysics applications, where the spatial discretization is usually performed using either central difference schemes [2], pseudo-spectral methods [3] or the finite element method [4]. [5] compared different spatial discretizations with higher order methods based on the spectral element method, p-FEM and isogeometric analysis for wave propagation problems. [6] demonstrated the advantages of isogeometric formulations compared to p-FEM with consistent mass matrices. However, this is requires the solution of a system of equations and thus drastically reduces the efficiency of explicit time integration schemes.

Finite difference models usually require a regular grid, which complicates the simulation of complex geometries. Specific extensions have been proposed to model heterogeneities [7] or nonuniform grids [8].

The spectral element method, originally introduced by [9], combines the Finite Element
Method with spectral methods on an element level. It is extensively used in the simulation of wave propagation, e.g. in seismology [10]. Similar to the Finite Element Method, the domain is decomposed into non-overlapping elements. Within each domain, the exact function is approximated by a spectral method. The interpolation is performed using Lagrange polynomials. The support points are chosen as the quadrature points of a Gauss-Lobatto rule. The position of the \( p \) nodes in a Gauss-Lobatto scheme are determined as the roots of the Legendre polynomial of order \( p - 1 \) plus the end points at \( \pm 1 \). The extension from 1D to 2D/3D is performed using a tensor product approach.

Standard Finite Element formulations often use an isoparametric approach. In the case of spectral elements, only the polynomial degree of the field variable is often increased, whereas the geometry interpolation is performed using a lower order basis (subparametric formulation).

One of the interesting features of the spectral element approach for wave propagation phenomena is the diagonal mass matrix, which substantially simplifies explicit time integration schemes. This is due to the evaluation of the element mass matrix \( M^e \) using the Gauss-Lobatto integration with the quadrature points being identical to the nodes. As a consequence,

\[
M^e_{ij} = \int_{\Omega} \rho \phi_i(\xi) \phi_j(\xi) \ J(\xi) \ d\Omega \approx \sum_{k=1}^{N} \rho w_k \phi_i(\xi_k) \phi_j(\xi_k) \ J(\xi_k),
\]

where \( J \) is the determinant of the Jacobian, \( \rho \) the density and \( \phi_j(\xi) \) the shape functions depending on the natural coordinates \( \xi \). The position of the quadrature points \( \xi_k \) and the corresponding weights \( w_k \) are determined by Gauss-Lobatto quadrature rules. In 2D and 3D, each diagonal entry has to be replaced by a block diagonal matrix. Assuming the determinant of the Jacobian to be constant (for regular elements), the polynomial order of the integrand in eq. (4) is twice the polynomial order of the shape functions. Using \( N \) nodes (1D), the integrand has degree \( 2N - 2 \). The Gauss-Lobatto quadrature is accurate for polynomials up to degree \( 2N - 3 \), i.e. the diagonal mass matrix is essentially the result of an under-integration, but the total mass is always conserved due to \( \sum_{i=1}^{N} w_i = 1 \). Furthermore, [11] showed that spectral convergence of the method is still maintained. The integration of the stiffness matrix using a Gauss-Lobatto quadrature is exact, since the integrand contains the derivatives of the shape functions with polynomial degree \( N - 1 \), i.e. the degree of the resulting polynomial is \( 2N - 2 \). Additionally, the stiffness matrix does not have to be assembled for explicit dynamics, but only its product with the displacement vector has to be calculated, which furthermore reduces the order of the polynomial to be integrated exactly to \( N - 1 \).

4 Time integration

In contrast to structural dynamics, where implicit solution schemes are often used, wave propagation problems are usually solved more efficiently with explicit methods. This is
due to the fact that the time step required for accuracy reasons is usually smaller than
the maximum time step related to stability and the computational costs for a single time
step with an implicit scheme are considerably higher due the necessity to solve a system
of equations.

Multiplying eq. (3) by the inverse mass matrix results in a modified system of equations
\[
\frac{\partial^2 \mathbf{d}}{\partial t^2} = M^{-1} (\mathbf{F}^c(t) - \mathbf{K} \mathbf{d}) = \mathbf{F}(t, \mathbf{d}).
\] (5)

This can only be efficiently solved for a diagonal mass matrix, which is automatically
fulfilled for spectral elements due to the under-integration of the mass matrix.

The simplest explicit method to solve eq. (5) is an explicit Euler method. An extension
to higher order methods is either obtained by using multi-step methods (e.g. Adams-
Bashforth methods), where multiple previous time steps are used to calculate the solution
at the current time, or single step methods with multiple stages such as the family of
Runge-Kutta methods. In this paper, the Velocity Verlet integration is used as a second
order method, since it is often implemented in commercial explicit dynamics codes such
as Ansys/LS-Dyna or Abaqus explicit. In addition, a fourth order Nyström method
developed for acoustic problems is extended to elastic wave propagation. Additionally,
other higher order methods have been investigated such as the classical fourth order
Runge-Kutta scheme or the 3/8 scheme [12], as well as higher order schemes with an
error control for adaptive step sizes. There was no significant difference between all
fourth order methods concerning the accuracy for a given step size. The Nyström method
was chosen, because it required slightly less computations.

4.1 Velocity Verlet

A popular method for the solution of eq. (5) is the Velocity Verlet integration [13]. The
displacements, velocities and accelerations for a time increment $\Delta t$ are computed as
\[
d(t + \Delta t) = d(t) + v(t) \Delta t + \frac{1}{2} a(t) \Delta t
\] (6)
\[
a(t + \Delta t) = \mathbf{F}(t + \Delta t, d(t + \Delta t))
\] (7)
\[
v(t + \Delta t) = v(t) + \frac{1}{2} (a(t) + a(t + \Delta t)) \Delta t.
\] (8)

4.2 Nyström method

Nyström methods are special solution procedures for second order differential equa-
tions, which are efficient for wave propagation problems in the case where damping can
be neglected [14]. In this paper, a fourth order symplectic method based on [15] is used.
Three intermediate accelerations are calculated

\[ V_1 = F(t + d_1 \Delta t, d(t) + d_1 \Delta t v(t)) \]  
\[ V_2 = F(t + d_2 \Delta t, d(t) + d_2 \Delta t v(t) + a_{21} \Delta t^2 V_1) \]  
\[ V_3 = F(t + d_3 \Delta t, d(t) + d_3 \Delta t v(t) + a_{31} \Delta t^2 V_1 + a_{32} \Delta t^2 V_2), \]

and the update of displacements and velocities is given by

\[ d(t + \Delta t) = d(t) + \Delta t v(t) + \Delta t^2 (\bar{b}_1 V_1 + \bar{b}_2 V_2 + \bar{b}_3 V_3) \]  
\[ v(t + \Delta t) = v(t) + \Delta t (b_1 V_1 + b_2 V_2 + b_3 V_3). \]

The coefficients are given as

\[ d_1 = \frac{3 + \sqrt{3}}{6}, \quad \bar{b}_1 = \frac{5 - 3\sqrt{3}}{24}, \quad b_1 = \frac{3 - 2\sqrt{3}}{12}, \quad a_{21} = \frac{2 - \sqrt{3}}{12} \]  
\[ d_2 = \frac{3 - \sqrt{3}}{6}, \quad \bar{b}_2 = \frac{3 + \sqrt{3}}{12}, \quad b_2 = \frac{1}{2}, \quad a_{31} = 0 \]  
\[ d_3 = \frac{3 + \sqrt{3}}{6}, \quad \bar{b}_3 = \frac{1 + \sqrt{3}}{24}, \quad b_3 = \frac{3 + 2\sqrt{3}}{12}, \quad a_{32} = \frac{\sqrt{3}}{6}. \]

### 4.3 Critical time step

The coupled system of ordinary differential equations in eq. (5) can be decoupled

\[ \frac{\partial^2 w}{\partial t^2} + \Lambda w = Q^{-1} M^{-1} F^e(t) \]

with the transformation and the eigendecomposition given by

\[ d = Qw, \quad M^{-1} K = Q \Lambda Q^{-1}, \quad \Lambda = \text{diag}(\lambda_1, ..., \lambda_n). \]

The critical time step of the full system is determined by the smallest critical time step of all decoupled equations. Any explicit time integration scheme leads to a mapping of the form

\[ y^{(t+1)} = G(y^{(t)}), \]

where \((t)\) characterizes the discrete time step. In our case, \(y\) is a vector containing displacements and velocities. Using a Taylor expansion, a small error \(\epsilon^{(t)}\) is propagated to the next time step by

\[ y^{(t+1)} + \epsilon^{(t+1)} = G(y^{(t)} + \epsilon^{(t)}) \approx G(y^{(t)}) + \frac{\partial G(y^{(t)})}{\partial y} \epsilon^{(t)}. \]

Combining eq. (19) and 20 leads to a first order approximation of the error in the time discretization

\[ \epsilon^{(t+1)} \approx \frac{\partial G(y^{(t)})}{\partial y} \epsilon^{(t)}. \]
For stability, a small error at time \( t \) should not be amplified at \( (t + 1) \), which corresponds to requiring \( \| \tilde{\lambda}_i \| < 1 \), where \( \tilde{\lambda}_i \) are the eigenvalues of the mapping \( \frac{\partial G(y^{(t)})}{\partial y} \).

For Velocity Verlet, this matrix is obtained by differentiation of eqs. (6) to (8) as

\[
\begin{bmatrix}
\frac{\partial d^{(t+1)}}{\partial d^{(t)}} & \frac{\partial d^{(t+1)}}{\partial v^{(t)}} \\
\frac{\partial v^{(t+1)}}{\partial d^{(t)}} & \frac{\partial v^{(t+1)}}{\partial v^{(t)}}
\end{bmatrix} = \begin{bmatrix}
-\frac{\theta}{2} + 1 & t \\
\theta^2 - \theta & t - \theta + 1
\end{bmatrix},
\]

with the abbreviation \( \theta = \lambda_i \Delta t^2 \). The eigenvalues are

\[
\tilde{\lambda}_{1,2} = 1 - \frac{\theta}{2} \pm \sqrt{\frac{\theta^2}{4} - \theta},
\]

which yields the final stability condition for the second order Velocity Verlet integration

\[
\Delta t_{\text{crit}}^{vv} \leq \frac{2}{\sqrt{\lambda_{\text{max}}}}
\]

with \( \lambda_{\text{max}} \) the largest eigenvalue in eq. (18). Similar results are obtained e.g. by [16].

For the Nyström method, [17] derived the critical time step using a spectral method for the spatial discretization for a problem of acoustic wave propagation. Extending the ideas derived for Velocity Verlet in the previous sections, the mapping is obtained as

\[
\begin{bmatrix}
\frac{\partial d^{(t+1)}}{\partial d^{(t)}} & \frac{\partial d^{(t+1)}}{\partial v^{(t)}} \\
\frac{\partial v^{(t+1)}}{\partial d^{(t)}} & \frac{\partial v^{(t+1)}}{\partial v^{(t)}}
\end{bmatrix} = \begin{bmatrix}
1 - \frac{1}{2} \theta + \frac{1}{24} \theta^2 - \frac{3 - \sqrt{3}}{728} \theta^3 \\
\frac{1}{\Delta t} \left( -\theta + \frac{1}{6} \theta^2 - \frac{1}{288} \theta^3 \right) & 1 - \frac{1}{2} \theta + \frac{1}{24} \theta^2 - \frac{3 + \sqrt{3}}{1728} \theta^3
\end{bmatrix},
\]

The eigenvalues are given by

\[
\tilde{\lambda}_{1,2} = \left( 1 - \frac{1}{2} \theta + \frac{1}{24} \theta^2 - \frac{1}{576} \theta^3 \right) \pm \sqrt{\left( 1 - \frac{1}{2} \theta + \frac{1}{24} \theta^2 - \frac{1}{576} \theta^3 \right)^2 - 1}
\]

with \( a = 1 - \frac{1}{2} \theta + \frac{1}{24} \theta^2 - \frac{1}{576} \theta^3 \). This results in the condition for stability of the fourth order Nyström method

\[
\Delta t_{\text{crit}}^{ny} \leq \frac{2\sqrt{2} - \sqrt{4 + \sqrt{2}}}{\sqrt{\lambda_{\text{max}}}} \approx \frac{2.5865}{\sqrt{\lambda_{\text{max}}}}.
\]

The computation of the eigendecomposition in eq. (18) is computationally expensive. An upper bound can be obtained using the largest element eigenvalue [18]

\[
\lambda_{\text{max}} < \max_{e=1..L} (\lambda^e),
\]

with \( \lambda_{\text{max}} \) the largest eigenvalue in eq. (18).
Jörg F. Unger

\[ p(t) \]

loading function

Ricker wavelet

\[ f=150 \text{ kHz} \]

\[ h=7\text{ cm} \]

\[ l=21\text{ cm} \]

Concrete with

- Young’s modulus \( E = 30000 \text{ MPa} \)
- Poisson ratio \( \nu = 0.0 \)
- Density \( \rho = 2400 \text{ kg/m}^3 \)

Figure 1: Geometry, loading function, material parameters and measurement points \( m_1, m_2, m_3 \) and \( m_4 \) of test example.

Figure 2: Displacement solution using quadratic elements integrated with fourth order Nyström using spectral elements (SEM) and finite elements with full and lumped mass matrices.

where \( L \) is the number of elements and \( \lambda^e \) is the maximum element eigenvalue of the elementary eigenvalue problem

\[ K^e d^e = M^e d^e \lambda \]  \hfill (30)

with the element mass and stiffness matrices \( K^e \) and \( M^e \). An alternative with a lower upper bound can be obtained using the Power iteration or the Lanczos method [19].

5 Numerical studies

The numerical studies have been performed for a very simple structure with a quasi one-dimensional wave propagation illustrated in fig. 1. For this purpose, the Poisson ratio has been set to zero, in order to avoid the excitation of a wave perpendicular to the longitudinal wave. For similar reasons, the excitation is not a point force (or a plate of a small diameter), but the whole boundary at \( (y = 0) \) is excited with a pressure function described by a Ricker wavelet with a peak frequency of 150kHz. This function is used by many authors to mimic acoustic emissions due to concrete cracking [20].

5.1 Spatial discretization

Figure 3 demonstrates the influence of using higher order spectral elements on the solution error. As opposed to finite element methods with a heuristic HRZ mass lumping [21], where increasing the order of the interpolation increases the error, the solution error for spectral elements decreases for higher order elements. In the example, quadrilateral elements on a regular element grid with an average node distance is 2mm are used, i.e. the size of the elements increases with a higher interpolation order. The time integration scheme was performed with the Nyström method presented in section 4.2 using a time step of \( 1.5 \times 10^{-7} \text{s} \). Even though the number of dofs in this example is identical for different
Figure 3: Influence of the spatial interpolation order for spectral elements with a constant node density and a fourth order time integration with a Nyström method.

orders of the interpolation, the computational effort for higher order elements is slightly reduced due to the decreased number of total integration points. This is due to the fact that integration points and nodes are at the same position, but nodes which are shared by multiple elements serve as an integration point for each of these elements.

5.2 Temporal discretization

The required spatial discretization for a given frequency resolution is often investigated in literature. The temporal discretization of the problem is rarely discussed and often second order accurate central difference methods such as the Velocity Verlet algorithm described in section 4.1 are used. The time step is usually chosen according to the stability criterion, which is calculated for Velocity Verlet according to eq. (24) or for the Nyström method from eq. (28). Two choices for the reference solution are investigated. At first, the temporal discretization of the semi-discrete problem in eq. (3) is investigated, i.e. the reference solution is obtained with a very short time step for a fixed discretization with linear or cubic elements. In fig. 4, the integrated error of the displacement solution is plotted for different time steps (1.0, 0.8, 0.5 and 0.1 times the critical timestep). The geometry is discretized with bilinear quadrilateral elements with an edge length of 1.175mm shown in fig. 4a, or with cubic elements with an edge length of 3.525mm shown in fig. 4b. In a second investigation, the results are compared with the reference solution (fine mesh and short time step), which is shown in fig. 4c for the case of linear elements and fig. 4d for cubic spectral elements.

All discretizations have 20 nodes per wavelength. The setup with linear elements in space with second order integration in time is a very common choice for the simulation of wave propagation problems with commercial software.

A spatial discretization with linear elements results in a rather large error, even if the temporal discretization is very fine. When using a fixed spatial discretization as in fig. 4a,
Velocity Verlet and Nyström converge to the same solution, but a very short time step for Velocity Verlet is required. Interestingly, the results are significantly different when comparing the results with the reference solution (fine discretization in space and time). In this case, the spatial error dominates. When using a higher order method such as the fourth order Nyström method, the error is almost independent of the time step. Using the second order Velocity Verlet integration, it is observed that the results obtained with the critical time step are almost as good as compared to shorter time steps or higher order methods. It can be concluded that for the current example discretized with linear elements it is sufficient to use a second order method, since the spatial discretization error is the dominant error term. The situation is fundamentally different when using higher order methods for the spatial discretization. Both reference cases shown in figs. 4c and 4d show the same trend. Based on these results it is concluded that higher order methods for the temporal discretization are only efficiently applied for a sufficiently accurate spatial discretization, i.e. higher order methods in space or a large number of low order elements. For cubic spectral elements in space combined with a fourth order Nyström method in
time, the critical time step, which is even larger compared to Velocity Verlet, can be used in the simulation to obtain a reasonably good approximation of the solution.

5.3 Computational efficiency

In the previous investigations, the error in the spatial or temporal discretization were compared independently. For the spatial discretization error, this was obtained using a very short time step with a fourth order time integration. For the temporal discretization, the error was calculated using a fixed spatial discretization. For practical problems, the total error is a combination of spatial and temporal error. Furthermore, a finer spatial discretization requires a shorter critical time step. In addition, higher order schemes are computationally more expensive. In order to evaluate the computational efficiency, the computation time for the different algorithms in order to obtain the same order of accuracy is compared.

In fig. 5 the integrated error is plotted over the computation time. The time discretization is performed using Velocity Verlet or Nyström with the critical time step calculated according to eqs. (24) and (28). The spatial discretization varies between 10, 20 and 40 nodes per wavelength. It is observed that for linear elements (order 1) the spatial discretization error dominates. Only for very fine meshes with more than 20 nodes per wavelength, higher order time integration schemes are computationally efficient. The situation is totally different for higher order spectral elements. In order to obtain the same order of accuracy, the computational effort of the second order Velocity Verlet algorithm is about three to four times larger compared to the time discretization with the fourth order Nyström method. Furthermore, the efficiency of a refinement characterized by the slope of the graphs is significantly higher for the fourth order scheme. Higher order time discretization schemes are only efficient, if the spatial discretization is sufficient to capture the wave number of interest. It is further noted that for the case of a time discretization with Velocity Verlet the performance of linear and quadratic spectral elements is almost identical. Only when using a higher order time integration schemes, the advantages of quadratic elements are visible.

It is further noted that higher order spectral elements with a second order method in time only lead to a moderate increase of the overall efficiency of the method. Only the combination of higher order methods for the spatial as well as the temporal domain leads to a significant performance boost.

Figure 5: Integrated error (over time) at the reflecting boundary $m_4$ as a function of computation time for spectral elements with order 2 and 4 and time integration with Velocity Verlet and Nyström.
6 Conclusions

The paper compares the efficiency of higher order methods for the simulation of ultrasonic wave propagation in elastic media. The spatial discretization is performed with spectral elements. It is shown that, in contrast to standard finite elements, higher order spectral elements can be efficiently used in explicit time integration schemes due to their diagonal mass matrix. For the temporal discretization, a fourth order Nyström method is compared with the second order Velocity Verlet integration. Numerical examples demonstrate that a discretization with higher order methods in space leads to a moderate performance boost. A real performance gain can only be obtained when combining higher methods in space and time simultaneously. When modelling complex geometries such as e.g. the heterogeneities in concrete on the mesoscale, the maximum size of the elements and thus the advantages of using large, higher order elements is limited. But it is demonstrated in the paper that using quadratic or cubic elements with a fourth order time integration scheme gives a huge performance gain compared to the linear elements with a second time integration.

References

68, 1999.


THE INFLUENCE OF GEOMETRICAL PARAMETERS ON THE PERFORMANCE OF PERFORATED BREAKWATER

RUBENS AUGUSTO AMARO JR\textsuperscript{1}, LIANG-YEE CHENG\textsuperscript{2} AND KAZUO NISHIMOTO\textsuperscript{3}

\textsuperscript{1}Numerical Offshore Tank (TPN-USP), Polytechnic School, University of São Paulo
Av. Prof. Mello Moraes, 2231, Cidade Universitária, 05508-030, São Paulo, SP, Brazil
rubens.amaro@usp.br

\textsuperscript{2}Department of Construction Engineering, Polytechnic School, University of São Paulo
Av. Prof. Almeida Prado, Trav. 2 No.83 Ed. Civil, Cid. Universitária, 05508-900 São Paulo, Brazil
cheng.yee@poli.usp.br

\textsuperscript{3}Numerical Offshore Tank (TPN-USP), Polytechnic School, University of São Paulo
Av. Prof. Mello Moraes, 2231, Cidade Universitária, 05508-030, São Paulo, SP, Brazil
knishimo@usp.br

Key words: Hydrodynamic impact, Breakwater, Simulation, Particle Method, MPS.

Abstract. Wave loads are of great concerns of coastal and offshore structures due to safety and operational issues, and lightweight protection devices onboard the floating structures have been proposed, such as vane type devices and perforated plates. In the present paper, a numerical study on the influences of the shape, number and size of holes on the performance of perforated lightweight breakwater, as well as the effects the spacing between the protection device and the protected installation are carried out. The simulations of the complex hydrodynamic phenomenon are carried out using a particle-based method. The results show that the impact loads on the perforated protection device depend mainly to the ratio between total hole and plate areas. On the other hand, the impact loads on the protected wall present nonlinear effects and the impulses in multiple hole configurations are slightly lower than that with single hole.

1 INTRODUCTION

Wave loads are of great concerns of coastal and offshore structures due to safety and operational issues. In order to mitigate the hydrodynamic impact loads, wave breakers or breakwaters have been used worldwide as protection devices. Nevertheless, the optimal design of the breakwaters by considering the wave conditions and the tolerable loads of the protected installations or equipment of the coastal and offshore structures still remains as a challenge mainly in case of the floating structures, where the weight of the protection devices might be considered in the design.

In the past, lightweight protection devices onboard the floating structures have been proposed, such as vane type devices and perforated plates \cite{1,2}. However, there are relatively few literatures regarding the performance, as well as the influences of the geometrical parameters, of the devices \cite{3,4} owing to complicated experimental settings and the limitation
of the traditional mesh based Computational Fluid Dynamics (CFD) approaches for the
modeling of the highly nonlinear hydrodynamic phenomena that involve large free-surface
deformation, wave breaking and relatively complex geometry of the protection devices.

In the present paper, in order to overcome the shortcomings of the traditional CFD
approaches, the influences of geometrical parameters on the performance of a breakwater are
investigated using a fully lagrangian meshless particle-based approach denominated Moving
Particle Semi-implicit (MPS) method [5], currently named Moving Particle Simulation. The
lightweight protection devices considered in the present study are perforated plates. For sake
of simplicity, the protected installation is modeled as a vertical wall. The incoming wave is
approximated by a collapsing of a water column subjected to gravity and hits first the
protection device, and later the vertical wall. At first, the convergence analysis is carried out
to optimize the numerical modeling. After that, the influence of protection device shape and
geometrical parameters, such as characteristic dimension of the holes, the ratio of perforated
area and spacing between the protection device and the protected wall, are investigated
through a series of numerical simulations. As a result, impact loads and impulses on the
protection device and protected wall are compared for different geometrical configurations of
the breakwater and possible correlations are discussed.

2 NUMERICAL METHOD

The governing equations of incompressible viscous flow are expressed by the conservation
laws of mass and momentum:

\[ \frac{Dp}{Dt} + \rho \nabla \cdot \vec{u} = 0 \]  
\[ \frac{D\vec{u}}{Dt} = -\frac{\nabla P}{\rho} + \nu \nabla^2 \vec{u} + \vec{f} \]

where \( \rho \) is the fluid density, \( \vec{u} \) is the velocity vector, \( P \) is the pressure, \( \nu \) is the kinematic
viscosity and \( \vec{f} \) is the external force vector.

In MPS method, the differential operators of the governing equations are replaced by
discrete differential operators on irregular nodes [6], which are derived from a model of
interaction between particles. For a given particle \( i \), the influence of a neighbor particle \( j \) is
defined by weight function \( \omega_{ij} \):

\[ \omega_{ij} = \begin{cases} \left( \frac{\|\vec{r}_{ij}\|}{r_e} - 1 \right)^2 & \|\vec{r}_{ij}\| \leq r_e \\ 0 & \|\vec{r}_{ij}\| > r_e \end{cases} \]

where \( r_e \) is the effective radius that limits the range of influence and \( \|\vec{r}_{ij}\| \) is the distance
between \( i \) and \( j \). In the present work, is used the effective radius \( r_e = 2.1l_0 \), where \( l_0 \) is the
initial distance between two adjacent particles.

The summation of the weight of all the particles in the neighborhood of the particle \( i \) is
defined as its particle number density \( n_i \), which is proportional to the fluid density:
For a scalar function $\phi$, the gradient and Laplacian operators are defined in Eq. 5 and Eq. 6, respectively:

\[
\nabla \phi = \frac{d}{n^0} \sum_{j=1}^{n} \frac{\phi_j - \phi_i}{\| \hat{r}_{ij} \|^2} \hat{r}_{ij} \omega_{ij}
\]

\[\nabla^2 \phi = \frac{2d}{\lambda_i n^0} \sum_{j=1}^{n} (\phi_j - \phi_i) \omega_{ij} \tag{6}\]

where $d$ is the number of spatial dimensions and $n^0$ is the initial value of $n_i$. Finally, $\lambda_i$ is a correction parameter so that the variance increase is equal to that of the analytical solution, and is calculated by:

\[\lambda_i = \frac{\sum_{j=1}^{n} \omega_{ij} \| \hat{r}_{ij} \|^2}{\sum_{j=1}^{n} \omega_{ij}} \tag{7}\]

To solve the incompressible viscous flow, a semi-implicit algorithm is used in the MPS method. At first, predictions of the particle’s velocity and position are carried out explicitly by using viscosity and external forces terms of the momentum conservation. Then the pressure of all particles is calculated by the Poisson equation for the pressure as follows:

\[\nabla^2 p^{t+\Delta t} - \frac{\rho}{\Delta t^2} \alpha p^{t+\Delta t} = k \frac{\rho}{\Delta t^2} \frac{n_i - n^0}{n^0} \tag{8}\]

where $n_i$ is the particle number density calculated based on the displacement of particles obtained in the prediction step, $\alpha$ is the coefficient of artificial compressibility and $k$ is the relaxation coefficient. Both $\alpha$ and $k$ are used to improve the stability of a computation method. Also, to provide the stabilization of MPS, the pressure gradient is modified as [7]:

\[\nabla p = \frac{d}{n^0} \sum_{j=1}^{n} \frac{P_j - \hat{\rho}_i}{\| \hat{r}_{ij} \|^2} \hat{r}_{ij} \omega_{ij} \tag{9}\]

where $\hat{\rho}_i$ is the minimum pressure between the neighborhood of the particle $i$.

Finally, the velocity of the particles is updated by using the pressure gradient term of the momentum conservation and the new positions of the particles are obtained.

In order to identify free surface particles, the particle number density is used as checking parameter. A particle is defined as free surface particle and its pressure is set to zero when its particle number density $n_i$ is smaller than $\beta_i n^0$. The value of $\beta_i$ used in this study is 0.85.

Solid wall boundary condition is imposed by using three layers of fixed particles. The particles that form the layer in contact to the fluid are denominated wall particles, of which the pressure is computed by solving Poisson equation for the pressure (Eq. 8), together with the fluid particles. The particles that form two other layers are denominated dummy particles. Dummy particles are used to assure the correct calculation of the particle number density of the wall particles. Pressure is not calculated in the dummy particles.
3 CASES OF STUDY

In the present study, the influence of geometrical parameters on the performance of a breakwater is investigated using a simplified configuration in which the protected installation is modeled as a rectangular vertical wall. For lightweight protection device, perforated plates with one, two or three holes are used. Both circular and square holes are considered. For sake of simplicity, the incoming wave is approximated by a collapsing of a water column subjected to gravity that hits first the protection device, and then the vertical wall, as shown in Figure 1. This configuration characterizes the theoretical dam break problem and has been proposed to study the green water flow on the deck [8]. The main dimensions of the breakwater model are given in Figure 2.

![Figure 1: Protection devices (green plates) with one (a), two (b) and three (c) holes, protected installation (brown wall) and water column at initial time.](image1)

![Figure 2: Dimensions of the breakwater model.](image2)

The simulation parameters for all the cases are presented in Table 1.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time step ((\Delta t))</td>
<td>0.0005 s</td>
</tr>
<tr>
<td>Particle distance (dp)</td>
<td>0.0025 m</td>
</tr>
<tr>
<td>Effective radius ((r_e))</td>
<td>2.1 dp</td>
</tr>
<tr>
<td>Free surface ((\beta_f))</td>
<td>0.85</td>
</tr>
<tr>
<td>Relaxation coefficient ((k))</td>
<td>0.75</td>
</tr>
<tr>
<td>Compressibility factor ((\alpha))</td>
<td>(10^{-8} \text{ m}^2/\text{kg})</td>
</tr>
<tr>
<td>Fluid density ((\rho))</td>
<td>1001 kg/m(^3)</td>
</tr>
<tr>
<td>Kinematic viscosity ((\nu))</td>
<td>(10^{-6} \text{ m}^2/\text{s})</td>
</tr>
</tbody>
</table>
4 RESULTS AND DISCUSSIONS

4.1 Convergence

The influence of distance of particle (dp) in the numerical model is verified for a case consisting of a protection device with a circular hole of 50 mm diameter (D). Four distances of particles are considered dp = 10 mm, 7.5 mm, 5 mm and 2.5 mm, which correspond to ratio dp/D = 0.2, 0.15, 0.1, 0.05, respectively. The force time histories on the protection device and the vertical wall are plotted for those distances of particles and given in Figure 3.

Figure 3 shows a similar behavior between the curves for all dp/D ratios, with exception of the pick force on the vertical wall that varies according to dp/D, where the maximum value of 17 N is obtained by dp/D = 0.05. For dp/D = 0.1 and 0.05 the oscillation of the numerical results reduced drastically when compared to those of dp/D = 0.15 and 0.2, indicating a more stable computation for the higher resolution models. Thus, in the following simulations, the distance of particle dp = 2.5 mm is adopted.

4.2 Breakwaters with circular or square holes

The influences of different shape, number and the characteristic dimension of holes, are investigated through a series of numerical simulations. Circular and square shapes, 1 to 3 holes and characteristic dimension ranging from 30 to 80 mm are taking into account. As main output, force time histories on the protection device and the vertical wall are plotted and analyzed. Finally, as a validation check and in order to verify possible correlations between the cases, impulses on the protection devices and vertical wall are analyzed for different ratio of perforated area. The geometrical dimensions of circular and square holes analyzed herein are summarized in Table 2.

Table 2: Geometry and dimensions of the holes of the breakwaters.

<table>
<thead>
<tr>
<th>Nº of holes</th>
<th>Circular diameter (mm)</th>
<th>Square length (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30, 40, 50, 60, 70, 80</td>
<td>30, 40, 50, 60, 70, 80</td>
</tr>
<tr>
<td>2</td>
<td>30, 40, 50, 60, 70, 80</td>
<td>30, 40, 50, 60, 70, 80</td>
</tr>
<tr>
<td>3</td>
<td>30, 40, 50</td>
<td>30, 40, 50</td>
</tr>
</tbody>
</table>
Figure 4: Force time histories on the protection device and vertical wall – One circular hole.

Figure 5: Force time histories on the protection device and vertical wall - One square hole.

Figure 4 provides the computed force time histories on the perforated protection devices with one circular hole and vertical wall. The time histories related to the protection devices with one square hole are given in Figure 5. It is important to point out here that the force on the protection device is computed on the weather side. The figures show that the increase of diameter or length, for the cases with one hole, leads to lower peak force in the protection device meanwhile the peak force on vertical walls are higher. Also, it is interesting to point out that the pattern of the force time histories on the perforated protection plate changes gradually from the small hole’s one impact peak pattern, which is close to that of a plain plate, to two peak pattern of larger holes. On the other hand, the behavior of the time history curves are independent to the hole shape, i.e., for both circular and square holes the curves show a similar pattern.

Figure 6 gives the snapshots of the free surface deformation of the collapsing water column obtained from the simulation of the perforated plate with one circular hole of 50 mm diameter.
Figure 6: Snapshots of the free surface deformation – perforated plate with one circular hole of 50 mm diameter.

Figure 7: Force time histories on the protection device and vertical wall – Two circular holes.

Figure 8: Force time histories on the protection device and vertical wall – Two square holes.

Figure 7 presents the computed force time histories on the protection devices and vertical wall for perforated plates with two circular holes and the time histories related to the protection devices with two square holes are illustrated in Figure 8. Similar to the breakwater with one hole cases, the figures show that the increase of diameter or length, for the perforated plates with two holes, leads to lower peak force on the protection device and higher peak force on the vertical wall, and the effects of the hole shape are also negligible. Considering the same characteristic dimensions, as the perforated areas of the square holes are larger than the circular ones, a higher momentum bypasses the breakwater with square holes.
Thus, when the water strikes the vertical wall, at about 0.25 s, the configuration with square holes present higher peak forces than that of circular holes.

Figure 9 gives the snapshots of the free surface deformation of the collapsing water column obtained from the simulation of the perforated plate with two circular holes of 80 mm diameter.

![Figure 9: Snapshots of the free surface deformation – perforated plate with two circular holes of 80 mm diameter.](image)

Figure 10 and Figure 11 give the force time histories associated to the protection devices.
with three circular or square holes, respectively. As noted in the previous cases, the computed time histories show the behaviors not very different from the previous cases with one or two holes.

Figure 12 gives the snapshots of the free surface deformation of the collapsing water column obtained from the simulation with three square holes of 40 mm length.

![Snapshots of the simulation with three square holes of 40 mm in length.](image)

In the present study, the impulses on the protection devices and the protected walls were computed considering the 2.0 s of simulation. For the protection devices, only the loads on the weather side are taking into account. Figure 13 show the computed impulse on the protection device and vertical wall as a function of ratio $h_t/p_t$, where $h_t$ is total hole area and $p_t$ is total plate area, for all the cases of perforated plate with circular holes. For the protection device with one, two and three holes, nearly linear and almost overlapped functions are obtained, showing that the magnitude of impulse depends only on the ratio $h_t/p_t$ and is inversely proportional to the ratio. On the other hand, the impulse on the protected wall, increases almost linearly with the ratio $h_t/p_t$ when the protection device has one hole. However, when the protection device has two or three holes, the curves present slight curvature. In addition to this, both curves of two and three holes are overlapped, but offset from the curve related to the protection device with one hole. These similarity between the curves of two and three holes and their difference from that of one hole clearly shows the non-linear behavior of the complex hydrodynamic generated by the flow separation and merging behind the perforated plate with multiple holes.

![Figure 13: Impulse vs. Ratio hole/plate area – Circle hole.](image)
The computed results of the relation between the impulses and ratio $h_t/p_t$, for the protection devices with square holes are given in Figure 14. From the figure, an extremely linear relation for the protection device indicates the impulse generated on the device decreases proportionally to the increase of ratio $h_t/p_t$, and depends exclusively on the ratio. For the impulse on the protected wall, the figure shows non-linear relation between impulse and the ratio $h_t/p_t$, as in the previous case. Nevertheless, slightly different from the cases with circular holes, while the curve of two hole is offset from that of one hole, the impulse in case of three holes is closed to that of two holes for lower $h_t/p_t$, and close to that of one hole when $h_t/p_t$ is higher than 0.3.

4.3 Spacing between the protection device and the protected wall

The effects of the spacing between the protection device and the protected wall are investigated by adopting four distances: 20 mm, 40 mm, 80 mm and 120 mm. In these cases, only protection device with two circular holes of 50 mm diameter are taking into account.

Figure 15: Force time histories on the protection device and vertical wall – The effects of spacing (protection devices with two circular holes of 50 mm diameter).

Figure 15 show the computed time histories of the force on the protection device and the vertical wall. From the numerical results, the impact force on the perforated protection device
is almost independent to the spacing. This is because the momentum transfer that occurs when the collapsing water column hits the protection device is independent to the position of the protected wall located behind the protection device. However, it is interesting to point out the delay of the small second impact force peak, which is associated to the second peak of the impact force on the protected vertical wall, when the spacing is creased.

On the other hand, the spacing greatly influences the impact force on the protected wall. The delay of the first impact on the wall due to enlargement of the spacing is clearly shown in the figure, as well as a slightly decrease in the magnitude of the initial impact force due to the deceleration of the incoming water. In all the spacing analyzed, after the water hits on the wall, the computed forces reach their peak value at about $t = 0.7$ s. However, the patterns of the time series change remarkably as the spacing varies. In the case with the smallest spacing of 20 mm, the mean impact force increases constantly, with oscillations that demonstrate violent fluid motions inside the gap between the protection device and the protected wall. At another extreme, the result of the case with the largest spacing shows that, after the initial contact, the impact force remains almost constant for about 0.4 s, and then an abrupt rise followed by a smooth decay of the force occurs. Also, smaller force peaks are computed at about $t = 1.0$ s. For intermediate spacing of 40 and 60 mm, the patterns of the force time histories present transition behaviors from the two extreme cases analyzed above.

Finally, Table 3 presents the impulse computed for the four spacing, during the 2.0 seconds of simulation. The increase of the spacing between the perforated protection device and protected wall leads to slightly lower impulse generated on the protection device. The computed impulse on the protected wall decreases remarkably with the increase of the spacing.

Table 3: Impulse as a function of the spacing between the perforated protection device and protected wall.

<table>
<thead>
<tr>
<th>Spacing [mm]</th>
<th>Impulse [N∙s]</th>
<th>Protection</th>
<th>Wall</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>19.10</td>
<td>16.07</td>
<td>35.17</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>18.88</td>
<td>15.10</td>
<td>33.98</td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>18.66</td>
<td>14.13</td>
<td>32.79</td>
<td></td>
</tr>
<tr>
<td>160</td>
<td>18.06</td>
<td>11.49</td>
<td>29.55</td>
<td></td>
</tr>
</tbody>
</table>

5 FINAL CONSIDERATIONS

In the present paper, the influences of the shape, number and the characteristic dimension of holes on the performance of perforated lightweight breakwater, as well as the effects the spacing between the protection device and the protected installation were investigated numerically. In order to model the highly nonlinear hydrodynamic phenomena that involve breaking waves and relatively complex shaped solid, a particle-based simulation method is adopted. For sake of simplicity, the protected installations were modeled as a vertical wall and the incoming wave is approximated by a collapsing of a water column subjected to gravity. Effects of the entrapped air are neglected. The computed results show that, for circular and square holes, the general patterns of the impact force time histories are similar regardless of circular or square hole shapes, as well as the number of the holes. Concerning the hole size,
for the cases with one hole, the increase of hole size leads to lower peak force in the protection device meanwhile the peak force on vertical walls are higher. The pattern of the force time histories on the perforated protection plate changes gradually from the small hole’s one impact peak pattern, which is close to that of a plain plate, to two peak pattern of larger holes. The results also show that the impulse on the protection device decreases linearly as the ratio between hole area and plate area increases, and it is independent to the hole shape, size and number. On the other hand, the impulse on the protected wall increases almost linearly with the ratio when the protection device has one hole. When the protection device has two or three holes, the curves present slight curvature and both curves of two and three holes are overlapped, but offset from the curve related to the protection device with one hole, showing the non-linear behavior of the complex hydrodynamics caused by the flow separation and merging behind the perforated plate with multiple holes. Finally, concerning the spacing between the protection device and the wall, the simulation results show that its main effect is on the pattern of the force and magnitude of impulse on the vertical wall.

6 AKNOWLEAGMENT

The authors thank Petrobras S.A. for the financial support during the development of the numerical simulator in Numerical Offshore Tank (TPN-USP).

REFERENCES

NUMERICAL SIMULATION OF VORTEX INTERACTIONS USING A FAST MULTIPOLe DISCRETE PARTICLE METHOD

RICCIARDI T. R.¹, BIMBATO A. M.² AND WOLF W. R.¹

¹Universidade Estadual de Campinas - UNICAMP
Campinas, SP, 13083-860, Brazil
tulioricci@fem.unicamp.br, wolf@fem.unicamp.br

²Universidade Estadual Paulista - FEG UNESP
Guaratinguetá, SP, 12516-410, Brazil
alexbimbato@feg.unesp.br

Abstract. The discrete vortex method (DVM) is based on a Lagrangian description of the vorticity transport equation. In order to numerically solve the DVM, one can split the vorticity equation into separate diffusive and convective effects. Several formulations can be used to model the diffusive effect, e.g. the random walk method, the core spreading method and the velocity diffusion method. The convection effect can be treated using the material derivative to avoid the solution of a non-linear term; this is the major advantage of the method since each discrete vortex is convected with the fluid velocity field. However, the solution of the fluid velocity field requires the contributions from the incident flow, the perturbation due to the body and the particle interactions. The latter contribution is computationally expensive since the Biot-Savart law is used to compute the induced velocity by all discrete vortices in the cloud. The fast multipole method is an attractive algorithm used to accelerate the expensive interactions of the discrete vortices. It reduces the computational cost of the Biot-Savart law from O(N²) to O(N), where N is the number of discrete vortices in the cloud for a particular time step. The present FMM algorithm is based on the original ideas of Greengard and Rokhlin, with modifications to further accelerate the solution. In the present work, both FMM and Biot-Savart law solutions are compared by calculating the vortex-vortex interactions for different cylinder wakes, previously generated by a DVM. The present numerical tool will be used in future computational simulations of aerodynamic flows past airfoils in pitching and plunging motions and vortex induced vibration problems.

Key words: Discrete Vortex Method, Fast Multipole Method, Fluid Mechanics.

1 INTRODUCTION

The discrete vortex method (DVM) is a numerical method based on a Lagrangian description of the vorticity transport equation, which is split into diffusive and convective effects [1]. Several formulations can be used to model the diffusive effect [1-4]. The convection effect can be treated using the material derivative to avoid the solution of a non-linear term and, therefore, each discrete vortex is convected with the fluid velocity field. The solution of the velocity field requires the contributions from the incident flow; the perturbation due to the body calculated by a panel method with appropriate boundary conditions; and the particle interactions calculated by the Biot-Savart law.
The DVM is a mesh-free method used to represent wakes and boundary layers of aerodynamic bodies by distributions of vortex particles. These wakes concentrate a large number of vortex particles behind the body and usually form several clusters of particles. The computation of interactions among particles requires an expensive convolution step which imposes a heavy limitation on the usage of the DVM for engineering problems. With that in mind, alternative ways are required to accelerate the DVM simulations.

The Fast Multipole Method is listed as one of the top 10 algorithms of the twentieth century (see Ref. [5]). It was developed by Greengard and Rokhlin [6] for the solution of N-body problems. The algorithm consists of clustering the influence of near-field elements into multipole expansions in order to evaluate their interaction at distant locations, i.e., the center of far away clusters. This way, the influence among different groups of particles is computed faster than the O(N²) operations required by the direct Biot-Savart law.

Due to the cost associated with vortex clustering, the reduction in time from O(N²) to O(N) is achieved when N is sufficiently large and a hierarchical multi-level multipole expansion is used. The method is not applicable among near-field clusters of vortex particles and, hence, these calculations are performed using the Biot-Savart law.

There are two different multi-level FMM algorithms (see details in Ref. [6] and [7]); the first one is suited when the distribution of particles is nearly uniform in the domain. In this case, the computational box is either uniformly refined while the number of particles inside any box is greater than an arbitrary value or it is uniformly refined to a prescribed level. The second implementation, which applies an adaptive refinement of the computational boxes, is highly efficient in cases where particles are clustered in specific locations. In the adaptive method, boxes are only refined in regions where clusters of particles exist.

As mentioned, wakes with several vortex particles concentrate behind the aerodynamic bodies, so one can argue that the adaptive method provides better cost reductions in this case since empty boxes are neither refined nor used. However, the non-stationary wake formed and the continuous creation of vortex particles along the body surface by the DVM requires an update of the pre-processing step and the continuous creation of lists of box interactions in the adaptive FMM every time step, considerably increasing the simulation time.

The natural clustering of vortex particles from the DVM causes high concentrations of particles in a small number of FMM boxes, which can make the analysis expensive due to the Biot-Savart interaction among adjacent boxes. It requires higher refinement levels in the FMM which, in turn, can increase the computational cost of the method due to the large number of computational boxes in the highest refinement level.

In this work, we propose to use the non-adaptive multi-level FMM with a single pre-processing step, along with several techniques which speed up both pre-processing and FMM steps in order to overcome the drawbacks mentioned. A further acceleration of the computational simulation is also obtained by the use of parallelization via OpenMP in several steps of the present FMM implementation. Numerical results obtained by the combined FMM-DVM are compared to those obtained by the direct DVM and a good agreement between the methods is shown. In this paper, we also investigate the effects of time step, refinement level and number of truncation terms in the series appearing in the FMM formulation on the numerical solutions. Simulations are performed using cylinder wakes obtained by the DVM at different time steps. These wakes are further convected and diffused using the Fast Multipole Method combined with the Discrete Vortex Method.
2 MATHEMATICAL MODEL

2.1 Discrete Vortex Method

The fluid flow is governed by the continuity equation

\[ \nabla \cdot \vec{u} = 0 \]  

(1)

and the unsteady 2D Navier-Stokes equations (N-S equations) expressed in terms of the vorticity vector, \( \vec{\omega} = (0,0,\omega) \):

\[ \frac{\partial \omega}{\partial t} + \vec{u} \cdot \nabla \omega = \frac{1}{Re} \nabla^2 \omega \]  

(2)

For the numerical simulation, the vorticity in the fluid flow is represented in a discrete form by a cloud of vortex blobs whose cores are filled with vorticity distributed according to a Gaussian curve. Since the left hand side of the N-S equations carries on the information about the convection and that the right hand side the information about the diffusion of vorticity, Chorin [1] proposes the use of the so called viscous splitting algorithm. According to this algorithm, at a time step of the numerical simulation, the convection process is governed by Eq (4), while the diffusion is governed by the Eq. (5):

\[ \frac{\partial \omega}{\partial t} + \vec{u} \cdot \nabla \omega = 0 \]  

(4)

\[ \frac{\partial \omega}{\partial t} = \frac{1}{Re} \nabla^2 \omega \]  

(5)

In this work the convection equation is written using a Lagrangian description and its solution is obtained using Euler’s method, while the solution of the diffusion equation is obtained using the random walk method. Once the vorticity field is modeled by a cloud of discrete vortices, the convection equation (4) is written in Lagrangian form as:

\[ \frac{d\vec{x}_j}{dt} = \vec{u}_j(\vec{x},t), \quad j = 1,NV \]  

(6)

where NV is the number of point vortices in the cloud and the velocity field \( \vec{u}(\vec{x},t) \) can be split in three parts according to Bimbato et al. [8]:

\[ \vec{u}(\vec{x},t) = \vec{u}_i(\vec{x},t) + \vec{u}_b(\vec{x},t) + \vec{u}_v(\vec{x},t) \]  

(7)

The contribution of the incident flow is represented by \( \vec{u}_i(\vec{x},t) \). For a dimensionless uniform incoming flow its components take the form:

\[ u_{i_1} = 1 \text{ and } u_{i_2} = 0 \]  

(8)

The body contribution, \( \vec{u}_b(\vec{x},t) \), can be obtained, for example, using the Boundary Element Method but it is not considered in this work.

The velocity \( \vec{u}_v(\vec{x},t) \), due to the vortex interactions, has its components written as:

\[ u_{v_i} = \sum_{j=1}^{NV} \lambda_j c_{ij} [x_k(t) - x_j(t)], \quad i = 1,2 \]  

(9)

where \( \lambda_j \) is the \( j \)-vortex strength and \( c_{ij} [x_k(t) - x_j(t)] \) is the \( i \) component of the velocity induced, at point \( x_k(t) \), by a unit strength vortex located at \( x_j(t) \).
The Euler method is applied to the solution of the Eq. (6) as:

\[ x_k^i(t + \Delta t) = x_k^i(t) + u_k^i(\vec{x}_k, t) \Delta t, \quad i = 1, 2 \]  

(10)

The solution to the diffusion, Eq. (5), in the Lagrangian context, is obtained using the Random Walk Method (see Ref. [9]) which aims to simulate the Brownian motion of the particles. According to this method, the random walk diffusion displacement \( \eta(x, t) \), with a zero mean and a \( (2\Delta t/Re) \) variance, is added after the convective step. Hence, the position of each vortex at the instant \( (t + \Delta t) \) is given by:

\[ x_k^i(t + \Delta t) = x_k^i(t) + u_k^i(\vec{x}_k, t) \Delta t + \eta_k^i, \quad i = 1, 2 \]  

(11)

The random displacement components are:

\[ \eta_k^1 = \frac{4\Delta t}{Re} \log \left( \frac{1}{p} \right) \cos(2\pi Q) \]  

(12)

\[ \eta_k^2 = \frac{4\Delta t}{Re} \log \left( \frac{1}{p} \right) \sin(2\pi Q) \]  

(13)

with \( P \) and \( Q \) being random numbers laying between zero and one.

2.2 Fast Multipole Method

The FMM was initially introduced by Greengard and Rokhlin [6], and optimized by Carrier et al [7]. Its main steps are: the creation of boxes in a pre-processing step (step 0), the clustering of particles into multipole expansions, calculated in the finest level of refinement, i.e., the smaller boxes, (step 1); the creation of multipole expansions at the parent (larger) boxes up to level 2 of refinement (step 2); the computation of multipole interactions to local representations among distant clusters from a list of “well-separated” boxes (step 3); the creation of local representations of the far-field induced by distant clusters of particles to children (smaller) boxes until the finest level of refinement is reached (step 4); and, finally, the computation of the total influence by all particles to a local particle inside a box at the highest refinement level of refinement (step 5). For more details on the FMM, see Ref. [10].

The main difference between the FMM proposed in Ref. [6] to that showed in Ref. [7] is how the boxes are created. In the first case, parent boxes at level \( l \) are always divided in 4 children boxes at level \( l+1 \), despite the fact that some boxes may be empty. In the second case, boxes containing clusters of particles may be refined in 4 children boxes, or may not, depending if the boxes are empty or not. Empty boxes are discarded and boxes with particles can be further refined into smaller boxes. In this case, the refinement is adaptive only in regions with vortex particles. Steps 1 and 2 in the FMM are known as upward pass and steps 3 to 5 are the downward pass. These steps use an expansion of Taylor's series, and the numerical accuracy is directly related to the number of truncation terms in the series.

As an example of the refinement in the computational box, Fig. 1 illustrates boxes at levels 3 and 5, the black and blue ones, respectively. As one can see, some boxes contain several particles (red dots), while others are empty. In this work, the empty boxes are ignored in order to reduce both computational time and memory cost. This is the main idea of the algorithm proposed by Carrier et al. [7], and, except by parent and children boxes, there is no interaction
of boxes at different levels, making the implementation easier than the adaptive FMM and faster than the non-adaptive one.

![Figure 1: Detail of an aerodynamic wake with vortex particles and the FMM boxes at different levels](image)

2.2.1 Pre-processing - step 0

The size of the initial box in the computational domain and the maximum refinement level are imposed in this work. The pre-processing steps, which are only performed once, create variables related to the boxes and are completely independent of the particles. Some of these variables are allocated using both the number of levels and the total number of boxes and, then, efficient ways to allocate the variables are implemented to avoid unnecessary memory allocation, e.g., derived data type.

Step 0, not shown by Greengard and Rokhlin [6] in their algorithm, determines all the particles inside a box and also the box for every particle. The quadrant where the particle is located can be determined with its relative position to the center of the initial box. Every box can be divided into 4 identical children boxes in level 1 of refinement, and, with the association of the quadrant to the child box, it is possible to determine which box contains the particle at level 1. This process can be repeated up to the penultimate level, resulting in the particle's finest level box.

Simultaneously, we create a list of all non-empty boxes and account of what particles are inside each of these boxes. This list permits an efficient memory allocation and also the reduction in analysis time in the loops of steps 1 to 4, described below.

Lists of well-separated boxes and their near neighbors are also determined in the pre-processing step. An important characteristic of the FMM refinement is that the boxes are similar to a regular structured mesh, so, it is easy to know the number of adjacent boxes. Denoting $B$ as the parent of a box $b$, $B$ can have up to 9 boxes sharing a node, being 8 near neighbors and $B$ itself. These 9 boxes have 36 children boxes, which can be classified as: $b$ itself (one box), the near neighbors of $b$ (8 boxes) or the 27 well-separated boxes in the interaction list of $b$. This way, it is possible to create both lists simultaneously.
2.2.1 Particle-to-multipole - step 1

The creation of multipole expansions is given by Eq. (14) via a Taylor's series truncated after $p$ terms, for $n$ particles with intensity $\lambda_i$ inside a box, with a complex distance of $z_i$ from the center of the box: $x$ and $y$ are the real and imaginary components of $z_i$, respectively (see Eq. (15)). The sum of all vortex particle intensities is given by (16). The error in FMM is strictly related to the truncation term, and the smaller the series, the larger the error. In this case, the series is a logarithmic function (vortex interaction), and for different applications, one should change the functions in the following equations.

$$a_{\text{finest level}} \equiv \sum_{k=1}^{p} \left( \sum_{i=1}^{n} (-\lambda_i) \left( \frac{z_i^k}{k} \right) \right)$$

(14)

$$z_i = x_i - x_{\text{box}} + j (y_i - y_{\text{box}})$$

(15)

$$Q \equiv \sum_{i=1}^{n} \lambda_i$$

(16)

In order to obtain faster analyses, the loop that creates the multipoles in the finest level can be performed only with the non-empty boxes, and not necessarily with all the boxes as initially presented by Greengard and Rokhlin [6].

2.2.2 Multipole-to-multipole - step 2

After the finest level multipole expansions $a(l)$ are created, the influences are translated to the center of the parent boxes at the level $l-1$, resulting in $a(l-1)$ in Eq. (17). These operations use the binomial coefficients in a Taylor series with $p$ terms. The intensity of the multipoles in level $l-1$, $Q(l-1)$ in Eq. (18), is the sum of the children intensities, $Q(l)$.

$$a(l-1) \equiv \sum_{i=1}^{4} \left( \sum_{k=1}^{p} \left( \sum_{kk=1}^{k} \left( a_i(l) z_{i-kk}^{k} \left( \frac{k - 1}{kk - 1} \right) - Q_l(l) \frac{z_{i-k}^k}{k} \right) \right) \right)$$

(17)

$$Q(l-1) \equiv \sum_{i=1}^{4} Q_i(l)$$

(18)

This step does not increase the error and, again, can be performed only in the non-empty boxes, from the finest level up to level 2.

2.2.3 Multipole-to-local - step 3

The multipole-to-multipole (M2M) steps are made up to level 2, which is the lowest level possible to have multipole-to-local calculations (M2L). The interaction lists of all boxes contain all the clusters that will interact via M2L to the box. An auxiliary variable is necessary to indicate if any box from the interaction list is empty or not, since the prep-processing steps in this work are independent from the particles.
The M2L step is given by Eq. (19):

\[ b(l) \equiv \sum_{j=1}^{n\text{box}} \left( \sum_{kk=1}^{p} \left( \sum_{i=0}^{p} \left( \frac{a_i(l)(-1)^{i} z_j^{kk+k-1}}{z_j^{kk}} \right) - \frac{Q_j(l)}{kk z_j^{kk}} \right) \right) \]  

(19)

where \( n\text{box} \) represents all the non-empty boxes from the interaction list. In this equation, \( z_j \) is the complex distance between the box with the multipole representation to that with the local representation, and \( a \) and \( Q \) are obtained from the previously calculated multipole expansion from box \( j \) in the interaction list of \( b(l) \).

2.2.4 Local-to-local - step 4

The local-to-local (L2L) step represents the translation of the influence coefficients from the boxes that are in the interaction list of the \( b \)'s parent box, \( B \), to a child box \( b \). This way, the influences of boxes at coarser levels than \( b \)'s level are evaluated in this step.

The influence of all distant boxes of \( b \) is the sum of the M2L steps from the boxes of its own interaction list and also from the L2L steps from its parent, \( B \). It is noticeable that there is no L2L calculation from level 1 to 2 since there is no M2L in level 1. For the same reason, there is not L2L in the finest level.

The L2L influence in a box in the level \( l+1 \) from its father at level \( l \) is given by (20)

\[ c(l+1) \equiv \sum_{i=1}^{4} \left( \sum_{k=0}^{p} \left( \sum_{kk=k}^{p} \left( b(l)(-z_i)^{kk+k-1} \right) \right) \right) \]  

(20)

Here, \( b(l) \) is the local representation of the far field multipole expansions at the parent box and \( z_i \) is the complex distance from the parent's center to its children.

2.2.5 Local-to-particle - step 5

When all contributions are evaluated in a box, they are translated by another Taylor’s series to all particles within the box. Hence, all the interactions from distant particles are evaluated with the steps M2M, M2L and L2L previously shown, while the near-field interactions are computed through the Biot-Savart law.

Greengard and Rokhlin [6] present the methodology to calculate the potential of the particles. In this work, we analytically compute the derivative of the induced particle to particle velocity through the potential. Equation (21) gives the translation from the center of a box to a particle, resulting in the induced velocity by the near field and distant particles:

\[ V_{\text{ind}} = V_{\text{nearfield}} + \sum_{k=1}^{p} b(l) k (z_i)^{k-1} \]  

(21)

In this equation, \( p \) is the truncation term in the Taylor series, \( b(l) \) is the sum of L2L and M2L in a box in the highest refinement level \( l \) and \( z_i \) is the complex distance from the particle \( i \) to the center of the box.

In the near-field interaction, we use a vortex with a Gaussian distribution of vorticity to avoid the singularity from a potential vortex and, to far-field calculations with the FMM, we use a simple potential vortex.
3 PRELIMINARY RESULTS AND DISCUSSION

The main interest in this work is to evaluate the time reduction and to perform an error analysis of the fast multipole method. In order to perform such assessment, FMM results are compared to those obtained by the Biot-Savart law. It is useful to remember that the diffusion process implemented in the current work is the Random Walk method, so the particles should never have the same position when comparing both methods. Since, in the present work, we are only focusing on the evaluation of the velocity field, including its error analysis and computational time, several subroutines from the DVM algorithm are not used, mainly the creation of new vortices, the effects of solid bodies and other more advanced features such as turbulence models.

3.1 Error analysis

We firstly investigate the error of the fast multipole method compared to the direct Biot-Savart law. Therefore, the diffusion process is ignored for this analysis, avoiding a random dispersion of the vortex. Hence, any difference is exclusively due to the numerical error.

Starting with a simple case, we compare the motion of two vortices with strength $\lambda = 0.01$, in an initial position equal to $\pm [0.5, 0.0]$. It is expected that they stay in a circular path due to their interaction on each other. The results from Biot-Savart law and FMM, after 10000 dimensionless time steps of 0.001 are shown in Fig. 2, where the green squares are the FMM solution obtained using 5 refinement levels and 30 terms in the Taylor series expansion. The red squares represent the solution obtained by the Biot-Savart law. The $L_\infty$ norm of the particle position is 5.55E-17 and the $L_2$ norm is 2.78E-17.

![Figure 2: Two identical vortices rotating in a circular path](image)
As shown in Eqs. (14) to (21), the method uses different levels and numbers of truncation terms in the Taylor series. To evaluate the error from FMM and DVM in the wake shown in Fig. 1 for a single time step, the $L_2$ and $L_\infty$ norms are obtained varying both parameters mentioned, ignoring the diffusion to avoid a random deviation.

**Table 1:** Results for the wake with 9000 vortex particles shown in Fig. 1.

<table>
<thead>
<tr>
<th>Truncation term</th>
<th>Norm</th>
<th>10</th>
<th>20</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$L_2$</td>
<td>$L_\infty$</td>
<td>$L_2$</td>
</tr>
<tr>
<td>LEVEL 4</td>
<td>4</td>
<td>4.84E-011</td>
<td>9.13E-008</td>
<td>5.47E-015</td>
</tr>
<tr>
<td>LEVEL 6</td>
<td>6</td>
<td>1.46E-010</td>
<td>2.87E-007</td>
<td>1.23E-014</td>
</tr>
<tr>
<td>LEVEL 8</td>
<td>8</td>
<td>1.73E-010</td>
<td>2.82E-007</td>
<td>2.11E-014</td>
</tr>
</tbody>
</table>

The error of one single particle will change the entire wake interaction for long time integrations. In other words, an initial small error can lead to larger total errors after several time steps, as seen in Fig. 3.

**Figure 3:** Comparison of Biot-Savart law and FMM after 1800 time steps

However, even with a growing error after 1800 time-steps of 0.01 dimensionless time units, the method keeps the main features of the wake, without compromising the physics of the problem. It is useful to remember that the analysis did not consider the random diffusion. If it is considered, the particle positions will diverge every time step, and thus, reducing the importance of the error from the FMM. One should keep in mind that the error of the time marching scheme also affects the overall accuracy of the solution.
3.2 Computational cost

The main interest in the FMM is its reduction from $O(N^2)$ operations to only $O(N)$. Several executions were performed with three different wakes, for refinement levels from 5 to 11, and with truncation term from 10 to 30. The executions used CPU parallelism up to 8 threads. The results in table 2 are the computational time, in seconds, to execute one single time step. A complete time analysis is a challenging task since several factors are involved simultaneously.

Table 2: Time in seconds for FMM analysis with different parameters

<table>
<thead>
<tr>
<th>Truncation</th>
<th>Particles</th>
<th>9000</th>
<th>112500</th>
<th>450000</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>10</td>
<td>0.212</td>
<td>0.230</td>
<td>0.233</td>
</tr>
<tr>
<td>E</td>
<td>6</td>
<td>0.127</td>
<td>0.135</td>
<td>0.145</td>
</tr>
<tr>
<td>V</td>
<td>8</td>
<td>0.075</td>
<td>0.090</td>
<td>0.161</td>
</tr>
<tr>
<td>E</td>
<td>9</td>
<td>0.082</td>
<td>0.187</td>
<td>0.372</td>
</tr>
<tr>
<td>L</td>
<td>10</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>L</td>
<td>11</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

It is possible to see that refining from level 8 to 9 in the 112500 wake, the time reduction from direct vortex interactions is greater than the increase in the multipole calculation for 10 terms in the series. However, for 30 terms, the multipole clustering computation is more expensive and it is not beneficial to further refine. This illustrates the non-linear characteristics of the FMM algorithm.

A direct calculation via Biot-Savart was performed for different wakes with serial and parallel calculations (using 4 and 8 threads). The results in table 3 show the time in seconds for each analysis.

Table 3: Time is seconds for DVM analysis

<table>
<thead>
<tr>
<th>Wake</th>
<th>9000</th>
<th>45000</th>
<th>112500</th>
<th>225000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serial</td>
<td>1.44</td>
<td>36.29</td>
<td>238.93</td>
<td>923.10</td>
</tr>
<tr>
<td>4 threads</td>
<td>0.40</td>
<td>19.12</td>
<td>122.95</td>
<td>438.48</td>
</tr>
<tr>
<td>8 threads</td>
<td>0.41</td>
<td>10.13</td>
<td>62.86</td>
<td>252.42</td>
</tr>
</tbody>
</table>

3.3 Time step influence

The Discrete Vortex Method requires a time marching method, or, in other words, another source of error. The use of Biot-Savart law is very expensive, making it difficult to use high-order time marching methods or high increments of time step. With the FMM, it is easier to use smaller time steps and high-order methods because of the faster simulations.

Figure 4 is obtained using the Euler explicit method, without considering the diffusion, for a dimensionless time of 2.5. It shows that the error due to the time marching is dominating over the accumulated FMM error. The wake used in the analysis has 225000 elements. It is noticeable that its numerical error can be much smaller than the time marching error. Hence, although a source of error, the FMM can reduce the global error due to the possibility of smaller time-steps.
From Figs. 3 and 4 it is possible to see a larger error, in terms of particle position, in the upstream region of the wake (where it was supposed to be the body). However, due to the absence of the body, this will be studied later when this contribution is present.
4 CONCLUSIONS

In this work we present an investigation of the numerical error and an analysis of computational cost of the Fast Multipole Method (FMM) applied to evaluate the vortex-vortex interaction in a cylinder wake. This method is compared to the Biot-Savart law, which is traditionally used by the Discrete Vortex Method to do such interaction. The FMM allows using smaller time-steps or high-order time-marching methods, (e.g. high-order explicit Runge-Kutta schemes) and thus reducing the global error of the flow simulation.

Future work will cover the investigation of turbulence models, different diffusion models, high-order time marching methods and the presence of the aerodynamic body, including the generation of vortices. The numerical tool under development will be applied for fast aerodynamic simulations of unsteady flows.

5 ACKNOWLEDGMENTS

The authors gratefully acknowledge the partial support for this research provided by Conselho Nacional de Desenvolvimento Científico e Tecnológico, CNPq, under the Research Grant No. 470695/2013-7. The work was also supported by Fundação de Amparo à Pesquisa do Estado de São Paulo, FAPESP, under the Research Grant No. 2013/03413-4. The computational resources provided by Centro Nacional de Processamento de Alto Desempenho, CENAPAD-SP, are also gratefully acknowledged. The first author is supported by a CAPES MSc. scholarship, which is also acknowledged.

REFERENCES

AN APPLICATION OF GENERALIZED DIFFERENCES TO UNSTEADY HEAT PROBLEMS SUBJECT TO MIXED BOUNDARY CONDITIONS

FRANCISCO J. DOMÍNUEZ-MOTA¹, J. GERARDO TINOCO-RUIZ, FERNANDO O. GUILLÉN-REYES, GERARDO TINOCO-GUERRERO AND ALEJANDRA VALENCIA-RAMÍREZ

¹Facultad de Ciencias Físico Matemáticas
Universidad Michoacana de San Nicolás de Hidalgo
Edificio “B”, Ciudad Universitaria
C.P. 58040, Morelia, México.
dmota@umich.mx

Key words: Generalized finite differences, Crank Nicolson, Heat Equation

Abstract. In this paper we present the application of a generalized finite difference Crank-Nicolson scheme to the numerical solution of the unsteady heat equation in 2 + 1 dimensions subject to mixed Dirichlet and Robin conditions, a problem which has not been extensively studied when the spatial domain has a non-rectangular shape. The generalized scheme is based on a second order difference scheme defined by an optimality condition, and it has been developed to solve Poisson-like equations whose domains are approximated by structured convex grids generated by the direct variational method. Numerical examples showing the accuracy of the method are presented.

1 Introduction

The unsteady heat equation subject to mixed Dirichlet and Robin conditions is the cornerstone of the heat and mass transfer problems. Its numerical solution is required in order to model a large number of systems in physics and engineering. However, the use of finite differences on non-rectangular spatial domains has not been extensively studied, in fact in general polygonal domains there are rather few reliable finite difference schemes which can be successfully applied.

In this paper we present a general difference scheme for the case where the domains of interest are simply connected polygonal domains which can be decomposed into rectangles or rings, due to the fact that the numerical solution of Poisson-like problems on the spatial domain via generalized finite differences, as discussed in [5], can be easily calculated.
2 Grid generation basics

A logically rectangular set

\[ G = \{ P_{i,j} | 1 \leq i \leq m, 1 \leq j \leq n \} \]

of points of the plane is a structured grid with quadrilateral elements for the polygonal \( \Omega \) of order \( m \times n \); a grid \( G \) is convex if and only if each one of the \( (m-1)(n-1) \) quadrilaterals (or cells) \( c_{i,j} \) of vertices \( \{ P_{i,j}, P_{i+1,j}, P_{i,j+1}, P_{i+1,j+1} \} \), \( 1 \leq i < m, 1 \leq j < n \), is convex and non-degenerate; the boundary which defines \( \Omega \) is simply the polygonal Jordan curve defined by the grid nodes \( \{ P_{1,j}, P_{n,j}, P_{i,1}, P_{i,m} | 1 \leq i \leq m, 1 \leq j \leq n \} \) (See Fig. 1).

3 General finite difference schemes for Laplace’s equation

A consistent general finite difference scheme at a grid point \( P_{i,j} = (x_{i,j}, y_{i,j}) \) for the cartesian laplacian operator

\[ Lu = u_{xx} + u_{yy} \]  

was presented, in [5], by considering the subgrid \( \{ P_{l,k} | i-1 \leq l \leq i+1, j-1 \leq k \leq j+1 \} \), for which it is required to find coefficients \( \Gamma_{l,k} \) such that the local truncation error satisfies
Under standard smoothness assumptions, the consistency condition up to second order

\[ [u_{xx} + u_{yy}]_0 - \sum_{i=-1}^{i+1} \sum_{k=-1}^{k+1} \Gamma_{l,k} u(P_{l,k}) = \]

\[ \left( -\sum_{l,k} \Gamma_{l,k} \right) u(P_0) + \left( -\sum_{l,k} \Gamma_{l,k} \Delta x_{l,k} \right) u_x(P_0) + \]

\[ \left( -\sum_{l,k} \Gamma_{l,k} \Delta y_{l,k} \right) u_y(P_0) + \left( 1 - \sum_{l,k} \frac{\Gamma_{l,k} (\Delta x_{l,k})^2}{2} \right) u_{xx}(P_0) + \]

\[ \left( -\sum_{l,k} \Gamma_{l,k} \Delta x_{l,k} \Delta y_{l,k} \right) u_{xy}(P_0) + \left( 1 - \sum_{l,k} \frac{\Gamma_{l,k} (\Delta y_{l,k})^2}{2} \right) u_{yy}(P_0) + \]

\[ O(\max\{\Delta x_{i,k}, \Delta y_{i,k}\})^3 \]

where

\[ \Delta x_{l,k} = x_{l,k} - x_{i,j}, \quad \Delta y_{l,k} = y_{l,k} - y_{i,j} \]

when

\[ i - 1 \leq l \leq i + 1, \quad j - 1 \leq k \leq j + 1. \]

In order to satisfy the consistency condition the \( \Gamma_{i,j} \) must be chosen so the coefficients of the partial derivatives of \( u \) up to second order vanishes. In general, these coefficients define a not well-determined linear system which must be solved in order to obtain the \( \Gamma_{l,k} \) values. Several approaches for this calculation can be used; a very efficient one is addressed in [2], where an heuristic scheme based on an unconstrained optimization problem which, due to the lack of restrictions, can be solved in a simple way: first, at every inner grid node we solve the least squares problem defined by the expansion coefficients

\[ \sum_{l,k} \Gamma_{l,k} \Delta x_{l,k} = 0 \]

\[ \sum_{l,k} \Gamma_{l,k} \Delta y_{l,k} = 0 \]

\[ \sum_{l,k} \Gamma_{l,k} (\Delta x_{l,k})^2 = 2 \]

\[ \sum_{l,k} \Gamma_{l,k} \Delta x_{l,k} \Delta y_{l,k} = 0 \]

\[ \sum_{l,k} \Gamma_{l,k} (\Delta y_{l,k})^2 = 2 \]
through the Cholesky factorization of its normal equations. Next, \( \Gamma_{i,j} \) is obtained from 
\[
\sum_{l,k} \Gamma_{l,k} = 0.
\]

4 Crank-Nicolson approximation

Once a scheme to approximate \( Lu \) has been calculated, what follows is straightforward: in order to solve the unsteady heat equation

\[
\frac{\partial u}{\partial t} = \nu (u_{xx} + u_{yy}),
\]

we use the difference scheme described in the previous section to approximate the right hand side of (2) (See [4]). This yields

\[
\frac{v^{(n+1)}_{i,j} - v^{(n)}_{i,j}}{\Delta t} = \frac{\nu}{2} \left( L_0(v^{(n+1)}_{i,j}) + L_0(v^{(n)}_{i,j}) \right),
\]

where \( v^{(n)}_{i,j} \) and \( L_0(v^{(n)}_{i,j}) \) are the approximations to \( u \) and \( u_{xx} + u_{yy} \) at the inner grid point \((n\Delta t, x_{i,j}, y_{i,j})\).

Solving iteratively the sparse system defined by (3) produces the approximation to the solution of (2).

5 Numerical test

For the first numerical test, we considered the polar grid

\[ G = \{ x_{i,j} = \rho_i \cos(\theta_j), y_{i,j} = \rho_i \sin(\theta_j) \} \]

with 81 points per side sketched in figure 1, where

\[ \theta_j = \frac{\pi}{2} - j \frac{\pi}{160}, \quad \rho_i = 0.75 + i \frac{0.25}{80}, \quad 0 \leq i, j \leq 80. \]

Notice that this grid is nearly orthogonal. The \( \Gamma_{i,j} \) coefficients were calculated for this grid. It must be noted that the calculation is very simple and required only for one inner grid node, since due to symmetry the coefficients are the same for every \( 3 \times 3 \) subgrid.

The exact solution for the test problem was selected as

\[
u(t, x, y) = e^{-2\pi^2 \nu t} \cos(\pi x) \cos(\pi y);
\]

for the test, Dirichlet conditions were considered at \( \theta = 0 \) and \( \rho = 0.75 \), and Neumann conditions for the normal derivative at \( \theta = \frac{\pi}{2} \) and \( \rho = 1 \). Neumann boundary condition was discretized on a six node subgrid at the boundary nodes as described in [5].

The numerical solution of equation (3) was calculated using 51 equispaced nodes for the
time interval \([0, 1]\); \(\nu = 0.25\) was chosen. The \(\| \cdot \|_2\) error norm for the test is summarized in table 1. It was calculated as the grid function
\[
\|u^{(n)} - U^{(n)}\|_2 = \sqrt{\sum_{i,j} \left( u_{i,j}^{(n)} - U_{i,j}^{(n)} \right)^2 A_{i,j}},
\]
where \(u^{(n)}\) and \(U^{(n)}\) are the exact and approximated solution calculated at the \(i, j\)th grid node and \(n\)th time level, respectively, and \(A_{i,j}\) is the area of the polygon defined by \(\{P_{i+1,j}, P_{i,j+1}, P_{i-1,j}, P_{i,j-1}\}\).

The calculated solution at \(T = 0.5\) is sketched in figure 2. In figure 3 a semilog plot shows the decrease of the quadratic error in the test as the time value increases.

6 Conclusions and future work

In this paper, we considered an application of the schemes addressed in [3] and [5] in order to calculate an approximated numerical solution of the classical heat equation using finite differences, and the results agree with published numerical solutions. It is worthwhile to notice that despite the non-rectangular shape of the spatial mesh grid used in the test, it was possible to generate a numerical approximation to \(Lu\), producing satisfactory results. In addition, it is clear that it can be extended to non-homogeneous diffusion equations in a easy way.
Figure 3: Logarithm of the quadratic error

REFERENCES


COMBINED FINITE ELEMENT AND MUSCULOSKELETAL MODELS FOR ANALYSIS OF PELVIS THROUGHOUT THE GAIT CYCLE

Emiliano P. Ravera\textsuperscript{1,2}, Marcos J. Crespo\textsuperscript{3}, Fabio A. Guarnieri\textsuperscript{4} and Ariel A. Braidot\textsuperscript{2}

\textsuperscript{1}National Council of Scientific and Technical Research (CONICET), Argentina. Provincial Route N° 11 km. 10, Oro Verde emilianoravera@bioingenieria.edu.ar

\textsuperscript{2}Laboratory of Biomechanics, FI-UNER, Argentina. Provincial Route N° 11 km. 10, Oro Verde abraidot@bioingenieria.edu.ar

\textsuperscript{3}Gait and Motion Analysis Laboratory, FLENI Institute for Neurological Research, Argentina. Escobar, B1625XAF mcrespo@fleni.org.ar

\textsuperscript{4}International Center of Computational Methods in Engineering, UNL-CONICET, Argentina. Provincial Route N° 168, Paraje el Pozo aguarni@santafe-conicet.gov.ar

Key words: Integrative Biomechanics, Gait, Musculoskeletal Model, FEM

Abstract. Biomechanical analysis has been performed widely by applying rigid-body musculoskeletal models, it enables to investigate the influence of muscle activation on body movements. In order to investigate the mechanics of the musculoskeletal tissues finite element (FE) method act as a complementary approach to in vivo and in vitro studies. For this reason, the musculoskeletal model and FE method are supplementary. Integrative biomechanics integrates biomechanics knowledge and methods at multiple scales to address fundamental and clinical problems at the tissue and organ level. Thus, an integrative approach is the natural approach to understand human pathophysiology. Although the utility of FE models in clinical practice are often criticized owing to their inadequate model verification and validation, so the interpretation as well as assessing the sensitivity of models are become vital.

The aim of the present study is therefore to develop a computational model of pelvis under the scope of integrative biomechanics. We include a rigid-body musculoskeletal model of the lower limb and we linked their estimation of joints movement and muscle forces with a FE model of the pelvis. An indirect validation of our model were done by
compare with a model similar to previous studies, and a sensitivity study throughout the gait cycle was conducted.

Our model shows the von Mises stress ranged on $0 - 12 \text{MPa}$ over of the surfaces of each model structure throughout the gait cycle. In addition, the inclusion of dynamic muscle forces throughout the gait cycle produces a smoothing in the peak stress areas but it not cause a diminishing in the maximum von Mises stress.

1 INTRODUCTION

One of the principal aim in many research laboratories as well as in medical and engineering applications are the mechanic and dynamic simulations of the neuromusculoskeletal system. The outcome of these process largely depends on the force-generation properties of the muscles, the anatomical features of the skeletal system and the underlying neuronal control system. In particular, the estimation of muscle forces by musculoskeletal models, usually requires of solving an optimization problem regardless if it is selected inverse or forward dynamics approach. So, biomechanical analysis of human locomotor system has been performed widely by applying rigid-body musculoskeletal models, it enables to investigate the influence of muscle activation on body movements [1].

In order to investigate the mechanics of the musculoskeletal tissues, mathematical tools, such as finite element (FE) method act as a complementary approach to in vivo and in vitro studies [2,3]. For this reason, the musculoskeletal model and FE method are supplementary.

Integrative biomechanics integrates biomechanics knowledge and methods at multiple scales to address fundamental and clinical problems at the tissue and organ level. Thus, an integrative approach is the natural approach to understand human pathophysiology. As a consequence, integrative biomechanics can identify important unsolved problems in basic biomechanics, and provide the route whereby their solution may be translated into advances in clinical medicine [4].

At this point, early studies of hip were focused on stress distributions within the femoral neck and the pelvic bones [5], where Anderson et al. (2005) it predicted values of peak stress to a range from 0-44 MPa during static loading [6]. Later works focused on predicting the stress distributions within the hip joint’s cartilage, where the cartilage region under the acetabular dome was identified as the peak site pressure. The values ranged from 4-8 MPa for given loads of 1500-3000 N [7–9]. On the other hand, FE models more similar to clinical practice, in a physiological manner, were shown for Phillips et al. (2007), Yoshida et al. (2006) and Abraham et al. (2013). The first research group developed a model of the pelvis that included muscular and ligament boundary conditions for the analyze of stress-strain field on a static load condition [10]; the second and third group analyzed the joint hip contact area and pressure in walking [11, 12].

Although the utility of FE models in clinical practice are often criticized owing to their
inadequate model verification and validation [13, 14]; Viceconti et al. (2005), proposed that biomechanics community should start to consider these instruments in this field accepting that numerical models can no be totally validated when are applied in simulating biological tissue. Consequently, the interpretation as well as assessing the models sensitivity are become vital [15].

The aim of the present study is therefore to develop a computational model of pelvis under the scope of integrative biomechanics. We include a rigid-body musculoskeletal model of the lower limb and we linked their estimation of joints movement and muscle forces with a FE model of the pelvis. The aim of the model is to examine whether the use of muscular boundary loads gives rise to altered stress distributions within the pelvis compare to those given by fixed boundary condition models. An indirect validation of our model were performed by comparing with a similar model of previous studies [5,6,10], and a sensitivity study under physiological load conditions and joint movement throughout the gait cycle was conducted.

2 METHODS

(a) Rigid-body musculoskeletal model  
(b) FE model of the pelvis

Figure 1: Integrative biomechanics approach of the lower limb.

2.1 Musculoskeletal Model

We used a musculoskeletal model with 23 degrees of freedom and 92 musculotendon actuators (Figure 1a). The degrees of freedom in the model included three translations and three rotations of the pelvis; three ball and socket joints, one located at the third lumbar vertebrae to model the interaction between pelvis and trunk and two at each hip joints. A custom joint with coupled translations and rotations at each knee, and a revolute joint at each ankle [16]. The model was scaled according to subject’s anthropometric measurements.

Inverse kinematics and inverse dynamics, from the experimental marker trajectories and ground reaction forces, were used to calculate joint angles and moments. The residual
reduction algorithm (RRA) was used to reduce residuals at the pelvis. RRA reduced the non-physiological forces and moments applied at the pelvis by making small changes to measurements that may have experimental error. Computed muscle control algorithm (CMC) [17] was used to estimate the muscle forces from the kinematics and to adjust the model obtained in RRA. Finally, we used the “Muscle Force Direction” plug-in to extract from musculoskeletal models useful information as direction and muscles patch to setting up a FE analysis of pelvis [18].

All steps through the dynamic musculoskeletal simulations were implemented by following the “guide to OpenSim work-flow and tools” and showed the values into a good threshold [19].

2.2 Geometric Definition of the Hip

High resolution CT image data (512 × 512, 320mm field of view, in-plane resolution 0.625 × 0.625mm, 0.33mm slice thickness) of a female cadaveric hip available in the U.S National Library of Medicine (www.nlm.nih.gov) was used for provided baseline geometry of the bones through the manual segmentation at the 3D slicer (www.slicer.org). Reconstructed surfaces were decimated to reduce the number of polygons and were filtered with Taubin Smooth Algorithm to remove segmentation artefacts using MeshLab (www.meshlab.sourceforge.net).

We assumed the hip joint to be a perfect ball and socket joint. We included a contact interface with a common radius for the femoral and acetabular cartilage was given by the midline of the joint space [20]. Harrington’s regression equations was used to predict the hip joint center location based on the images [21]. So, the pelvic bone was established in the neutral kinematic position [22] and femur was moved following the hip joint angle of the sagittal plane during the gait cycle.

Triangular surface mesh for cortical bones and cartilages was converted to a linear 3-node shell element mesh. The interiors of cortical shell meshes were filled with tetrahedral elements, i.e., volumetric meshes were generated to represent each FE model structure using Gmsh [23].

2.3 Material Properties and Boundary Condition

Pelvic bone was assigned to behave as a homogeneous, isotropic, linear-elastic material, with an elastic modulus $E = 7GPa$ and Possion’s ratio $\nu = 0.3$ which effectively describe the weighed stiffness properties of the cortical and cancellous bone components [24]. Femur bone was modeled as a rigid body structure; and cartilages were modeled as a homogeneous, isotropic, nearly incompressible, neo-Hookean hyperelastic material with shear modulus $G = 13.6MPa$ [7].

A rigid interface was used to define the interface between femoral cartilage and femur bone. Two sliding contact algorithms based on the augmented Lagrangian approach were used to define acetabular cartilage to pelvic bone and cartilages interfaces, respectively.
Frictionless contact was assumed for the cartilage-to-cartilage interface, since the friction coefficient between cartilage surfaces is low (0.01-0.02 in the presence of synovial fluid) [8]. Application of fixed boundary conditions at the sacro-iliac joints is seen to be consistent with load transfer occurring between the pelvic bones and the sacrum in these regions. So, nodes along the sacro-iliac joint and the pubis symphysis joint were constrained as fixed to simulate these interfaces [6, 10].

To achieve a simulation of the stress that affect the hip in walking three conditions were included:

- Rotation of the hip joint center: a flexion-extension hip joint.
- Load compression were prescribed: a vertical net force of knee joint.
- Muscle Forces: fifteen vectorial muscle forces of the main muscles with origin or insertion patch in the pelvic bone.

We used the spot found by the OpenSim plug-in [18] and nodes around it were estimated the origin or insertion muscle patches in the FE model of the pelvis, see Figure 1b.

2.4 Convergence Studies

Mesh convergence studies are fairly prevalent in the biomechanics literature. Convergence criterion used in this work refers only over the quantification hypothesis, without considered other numerical errors. FE analysis was converged if the size of the elements decreased, and therefore the number of nodes and elements increased, so, the numerical solution tends to the exact solution. Mesh convergence studies do not guarantee that model predictions are accurate. Rather, they ensure that a finer discretization would likely not change the predictions significantly [13]. The number of tetrahedral elements was increased, so the stress variation of FE model was analyzed.

2.5 Sensitivity Studies

Changes in material properties assumed were performed to investigated how such properties affects the stress predictions in physiological conditions throughout the gait cycle. The baseline cartilages shear modulus were altered by $\pm 1SD$ using standard deviations for human cartilage [25]. To ascertain the effects of the assumption of cartilage incompressibility, bulk to shear modulus ratios of 100:1 ($\nu = 0.495$) and 10:1 ($\nu = 0.452$) were analyzed [7]. To quantify the effect of bone modeled we analyze the variation in the FE model prediction when the elastic modulus was changed in $\pm 10\%$ of its experimental value [24].

2.6 Data Analysis

The 3D model was imported to FE solver FEBio 1.7 for nonlinear, large deformation stress-strain analysis [26]. Also PreView and PostView was used to pre and post-process
Table 1: Maximum von Mises stress of mesh convergence study in MPa.

<table>
<thead>
<tr>
<th>Structure</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetabular Cartilage</td>
<td>2.695</td>
<td>8.463</td>
<td>8.465</td>
</tr>
<tr>
<td>Femoral Cartilage</td>
<td>2.695</td>
<td>4.441</td>
<td>4.539</td>
</tr>
<tr>
<td>Pelvis (*)</td>
<td>0.256</td>
<td>0.464</td>
<td>0.488</td>
</tr>
</tbody>
</table>

(*) Average of von Mises stress.

of FE model (www.febio.org). Structural mechanics quasi-static analysis and a line search method to improve the convergence of the nonlinear quasi-Newton BFGS solution algorithm was used. A static load condition was assumed to study the mesh convergence. The density of each mesh was adjusted until it doesn’t presented change in the stress average above of 5%. On the other hand, for the sensitivity studies, changes in peak pressure were reported as results of the complete dynamic loading scenario analyzed. So, we study a total of twenty four models before reaching the optimal configuration. In this optimal configuration of the FE model, displacement and von Mises stress distribution throughout the gait cycle was analyzed.

3 RESULTS AND DISCUSSIONS

For acetabular cartilage, femoral cartilage and pelvic bone three models of each were analyzed to determine the optimal density elements of each volumetric meshes upon a static load of one body weight (BW). Model 1 for acetabular cartilage have 5,000, Model 2 have 9,579 and Model 3 have 18,405 elements. For femoral cartilage the three models have 4,705, 22,336 and 130,156 elements respectively; and for pelvic bone these are 27,276, 94,757 and 218,208 elements. Increasing density mesh over Model 2 for each structures produces an improvement in the prediction of the stress field below of the 5% (see Table 1), proving a convergence of meshes. Therefore the final FE meshes consist of $\sim$137,000 tetrahedral elements (acetabular cartilage = 9,579 elements, femoral cartilage = 22,336 elements, pelvic bone = 94,757 elements and femur bone = 10,620 elements).

Figure 2 shows the outcome of the musculoskeletal model that is used as input of our pelvis FE model. Displacement and compression prescribed represent the hip joint movement in the sagittal plane and the vertical net force of knee joint respectively; the muscle forces are the modulus of the vector forces from the muscles that present their origin or insertion patch in the pelvic bone. All forces are normalized respect to subject body weight.

Figure 3 shows the maximum von Mises stress achieved throughout the gait cycle for different values of the model material properties. We analysed a sensitivity study for two different FE model of the pelvis. A model without muscle forces and similar fixed boundary conditions of the model proposed by Anderson et al. [6]; and in the other hand a new model with similar boundary conditions but with the inclusion of vectorial muscle
forces throughout the gait cycle. At this time the acetabular cartilage and pelvic bone of the FE model without muscle forces are not significantly sensitive to material properties variations. However we shown that when the muscle forces are included, acetabular cartilage and pelvic bone behaviors are more sensitive to material properties changing; mainly in the assumption of cartilage incompressibility.

The overall behavior of the FE models is illustrated in Figure 4, that shows the displacement of the pelvic bones along the sagittal plane. It is observed that the lateral part of the ilium and the ischium move in the anterior direction, but the sacro-iliac joint and the pubis remains stationary due to the applied boundary conditions. It is observed that the magnitude of displacement around of $0 - 0.1\text{mm}$, in concordance with the model proposed by Phillips et al. [10], and the inclusion of the muscle forces produce a smoothing of the displacement zones by showing a more physiological behavior throughout the gait cycle.

Anderson et al., shown that the von Mises stress ranged was between $0 - 44\text{MPa}$ for a subject-specific hip model but this peak appears to be unrealistic, since bone under
such high repetitive stress could probably degenerate. On the other hand, it is likely that a more physiological load could reduce this peak [6]. In concordance with this, using a physiological load as vertical net force of knee joint when the hip joint moves throughout the gait cycle, our model shows the von Mises stress ranged between 0−12 MPa (see Figure 3). This improves highly the stress field. Figure 5 shows how the stress are transferred over the surfaces of each model structure throughout the gait cycle. Maximum peak happens in sacro-iliac joint in the antero medial area as a response to the weight load around the 10% of the gait cycle, nevertheless this peak moves to posterior-lateral zone when the hip joint is between 50−60% of the gait cycle. In addition, the inclusion of dynamic muscle forces throughout the gait cycle produces a smoothing in the peak stress areas (as it was presented by Phillips et al. [10]) but it doesn’t cause a diminish in the maximum von Mises stress. So, it leads to avoid a point stress concentration in the pelvic bone and we think that the FE model with muscle forces represents a more realistic model approach to the physiological behavior of the lower limb during walking.

4 CONCLUSIONS

Integrative biomechanics can be used to guide therapies for chronic conditions such as osteoarthritis, osteoporosis and cerebral palsy; as well as computer-aided surgical planning and robotics-assisted surgery [4]. In particular, the aim of the present study was to examine the use of muscular boundary loads in the pelvis models with a more physiological behavior.

However, direct comparison of displacement and stress values of these models with
Figure 4: Anteriorposterior displacement of the FE models of the pelvis. Displacements in the model with muscle forces are found to be more consistent with physiological observations.

prior experimental studies is difficult because of the differences in loading modalities and measurement techniques. For this reason we proposed a development and evaluation of a computational model throughout the gait cycle with a complete set of open-source software, believing that the biomechanics community will be benefit with a greater degree of collaboration and peer review.

In conclusion, this paper presents a new model that have a supplementary approach between a rigid-body musculoskeletal model and FE method. In addition the model includes physiological load conditions and movement (like muscle forces, knee reaction force and hip joint movement) to find a more realistic stress field.

ACKNOWLEDGEMENTS

The authors thank the FLENI Institute for Neurological Research for providing data from healthy subjects and the National Council of Scientific and Technical Research and PICTO 222-2009 (AGENCIA) for providing the funds needed for this research.
a) FE model without muscle forces.

b) FE model with muscle forces.

Figure 5: Von Mises stresses of pelvic bone throughout the gait cycle in both models with and without muscle forces.

REFERENCES


Dalstra M., Huiskes R., and van Erning L. Development and validation of a three-
dimensional finite element model of the pelvic bone. *Journal of Biomechanical En-

Anderson A.E., Peters C.L., Tuttle B.D., and Weiss J. Subject-Specific Finite Ele-

Anderson A.E., Ellis B., Maas S., Peters C., and Weiss J. Validation of finite el-

Harris M., Anderson A.E., Henak C., Ellis B., Peters C.L., and Weiss J. Finite 


Yoshida H., Faust A., Wilckens J., Kitagawa M., Fetto J., and Chao E.Y. Three-
dimensional dynamic hip contact area and pressure distribution during activities of 

Abraham C.L., Maas S., Weiss J., Ellis B.J., Peters C.L., and Anderson A.E. A new 


MECHANICAL PROPERTIES OF FOOT TENDONS – IN VITRO STUDY

E. MORALES-ORCAJO1, 2, R. BECERRO-DE-BENGOA-VALLEJO3, M. LOSA-IGLESIAS4, J. BAYOD1

1Group of Structural Mechanics and Materials Modeling (GEMM) Aragon Institute of Engineering Research (I3A). University of Zaragoza, Zaragoza, Spain. Biomedical Research Networking Center in Bioengineering, Biomaterials & Nanomedicine Campus Río Ebro, C/ María de Luna s/n, Edificio Betancourt50018, Zaragoza, Spain Email: enriquem@unizar.es & jbayod@unizar.es

2Group of Biomechanics Engineering UFMG - (MecBio) School of Engineering, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil Webpage: http://lattes.cnpq.br/6122905565172564

3Nursing Physiotherapy and Podiatry Faculty. Medicine Faculty, Complutense University, Madrid, Spain

4Faculty of Health Sciences, Rey Juan Carlos University, Madrid, Spain

Key words: Foot tendons; Mechanical properties; Mechanical testing; Tissue characterization

Abstract. The aim of the present study was to determine the mechanical properties of main intrinsic and extrinsic human foot tendons in vitro. Seventy six uniaxial tensile tests were performed to obtain the stress-strain curve of different foot tendons. The specimens were harvested from four frozen cadaver feet and included: Extensor and Flexor tendons of all toes, Tibialis Anterior and Posterior tendons, Peroneus Brevis and Longus tendons. Young’s modulus was calculated for each sample. Ultimate tensile stress and strain failure were also recorded. Furthermore, the characteristics of the stress-strain curve, initial strain and linear region, were quantified. All tests showed the characteristic tendon stress-strain curve with an initial strain without load followed by a linear region until macroscopic failure. The mean strain failure was 13 ±6.2%. The Young’s modulus of all samples ranged from 0.1 to 0.8GPa with a mean value of 0.34 ±0.18GPa. Two different tendon behaviors could be distinguished. Tibialis and Peroneus tendons exhibited higher values of strain failure if compared to Flexor and Extensor tendons which had higher Young’s modulus.
1 INTRODUCTION

Within foot soft tissue, tendon is the tissue with less information available. Recent foot computational studies incorporate approximate representations of tendons [1,2], thus, deeper study of foot tendons properties has become necessary. Mechanical properties of foot tendons have been studied previously, especially the Achilles tendon properties, which is one of the most important tendons in the human body. Young’s modulus of Achilles tendon from in vitro tensile test are in the range of 0.5-0.8GPa [3–5]. The modulus estimated from in vivo studies are between 0.8 to 2 GPa [6–8]. However, information reported about other foot tendons is scarce and incomplete. Only material properties of Extensor Digitorum Longus (EDL) and Tibialis Anterior (TA) have been reported [9,10].

The objective of the present study is to determine the material properties of foot tendons involved in body stability. Information on the mechanical properties of these lesser-studied tendons is very useful for computational simulation but also for clinical applications, such as tendon repairs.

2 MATERIAL AND METHODS

Tendons included in the experiments were sorted in two groups: the long tendons involved in flexion and extension of the toes, on the one hand and the thick tendons intervening on the inversion and eversion of the ankle, on the other hand. The former includes the Extensor Digitorum Brevis (EDB) and the EDL which extent lesser toes, the Extensor Hallucis Longus (EHL) which extends the great toe, the Flexor Digitorum Brevis (FDB) and the Flexor Digitorum Longus (FDL) which flex the four lateral toes and the Flexor Hallucis Longus (FHL) which flexes the hallux. The latter involves the TA and the Tibialis Posterior (TP) which invert the foot, and the Peroneus Brevis (PB) and Peroneus Longus (PL) which evert the ankle. Those muscles and, by extension, their tendons are important because they enable us to stay balanced in upright position.

2.1 Specimens preparation

A total of 76 tendons samples were taken from four male elder donors. A sample of each tendon was cut from the most relative uniform cross-sectional area (CSA) and all soft tissue around the tendon was removed. After the dissection, the samples were frozen and kept at a temperature of -20ºC and gradually thawed 12h before the experimental tests. Specimens were kept hydrated until the time of testing at room temperature (~22ºC).

The tendon CSA was approximated by an ellipse and measured in three different positions along the sample with a caliper. Hence, two measurements, width and height, were used to calculate the CSA. The minimum of the three measurements was selected.

2.2 Mechanical testing

Uniaxial tensile test were performed to characterize tendon behavior (Instron Ltd., U.K., model 5548) (Figure 1). Each tendon was initially stretched until testing machine recorded a force of 0.01N to remove any slack. The specimens were then subjected to a tensile force at a displacement rate of 0.1 mm per second. Each specimen was loaded to failure. Strain was measuring using clamp-to-clamp displacement. The trials with evidences of slipping were discarded.
3 RESULTS

3.1 Cross sectional area

The tendon specimens showed relatively uniform CSA along the long axis. Means, standard deviations and ranges of CSA of the tendons tested are summarized in Table 1. Ranges are presented in order to facilitate comparison with other works. The mean CSA of digitorum tendons was much smaller than hallucis tendons and inversion-eversion tendons.

Table 1: Means and standard deviation of cross-sectional areas in mm$^2$.

<table>
<thead>
<tr>
<th>Tendon</th>
<th>Mean</th>
<th>±SD</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDB</td>
<td>3,22</td>
<td>0,84</td>
<td>1,53 - 4,56</td>
</tr>
<tr>
<td>EDL</td>
<td>4,85</td>
<td>0,91</td>
<td>3,30 - 6,01</td>
</tr>
<tr>
<td>FDB</td>
<td>2,59</td>
<td>1,04</td>
<td>1,48 - 5,18</td>
</tr>
<tr>
<td>FDL</td>
<td>5,29</td>
<td>2,04</td>
<td>1,54 - 9,03</td>
</tr>
<tr>
<td>EHL</td>
<td>8,47</td>
<td>1,55</td>
<td>6,91 - 10,41</td>
</tr>
<tr>
<td>FHL</td>
<td>16,23</td>
<td>2,81</td>
<td>12,53 - 18,69</td>
</tr>
<tr>
<td>TA</td>
<td>26,98</td>
<td>2,93</td>
<td>23,75 - 29,85</td>
</tr>
<tr>
<td>TP</td>
<td>24,77</td>
<td>5,88</td>
<td>20,29 - 33,18</td>
</tr>
<tr>
<td>PB</td>
<td>11,91</td>
<td>2,32</td>
<td>10,13 - 15,22</td>
</tr>
<tr>
<td>PL</td>
<td>17,03</td>
<td>3,73</td>
<td>13,22 - 21,17</td>
</tr>
</tbody>
</table>
Table 2: Strain of each region of the stress-strain curve.

<table>
<thead>
<tr>
<th></th>
<th>EDB</th>
<th>EDL+EHL</th>
<th>FDB</th>
<th>FDL+FHL</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Strain</td>
<td>1.8%</td>
<td>2.5%</td>
<td>2.2%</td>
<td>3.2%</td>
<td>2.5%</td>
</tr>
<tr>
<td>Toe region</td>
<td>4.0%</td>
<td>5.2%</td>
<td>5.0%</td>
<td>6.1%</td>
<td>5.2%</td>
</tr>
<tr>
<td>Yield point</td>
<td>7.0%</td>
<td>9.3%</td>
<td>8.0%</td>
<td>9.4%</td>
<td>8.6%</td>
</tr>
<tr>
<td>Failure point</td>
<td>8.6%</td>
<td>11.5%</td>
<td>10.2%</td>
<td>12.4%</td>
<td>10.9%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>TA</th>
<th>TP</th>
<th>PB</th>
<th>PL</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Strain</td>
<td>6.3%</td>
<td>7.1%</td>
<td>2.2%</td>
<td>4.2%</td>
<td>4.9%</td>
</tr>
<tr>
<td>Toe region</td>
<td>12.9%</td>
<td>16.2%</td>
<td>5.2%</td>
<td>8.8%</td>
<td>10.8%</td>
</tr>
<tr>
<td>Yield point</td>
<td>18.7%</td>
<td>22.6%</td>
<td>11.7%</td>
<td>13.6%</td>
<td>16.7%</td>
</tr>
<tr>
<td>Failure point</td>
<td>23.3%</td>
<td>26.1%</td>
<td>18.3%</td>
<td>17.4%</td>
<td>21.3%</td>
</tr>
</tbody>
</table>

3.2 Stress - strain curve

All trials had the characteristic stress-strain curve of tendon tissue. Three distinct regions were identified: the toe region characterized by a non-linear strain with small tension, the linear region where the stress increase linearly with the strain and the yield region or the region where starts the macroscopic failure. The toe region, in turn, is possible to divide in two parts: the considerable initial strain and the non-linear behavior.

Taking as a reference the Young’s modulus of each test, we quantified those regions. Hence, the initial strain was defined as the part of the curve from its beginning to a deviation greater than 20% of the Young’s modulus; the linear region was defined as the part of the curve with the curvature less than 20% of the Young’s modulus. The regions resulting visually fit with the regions description. The values of those regions are shown in Table 2. The average was weighted with the number of tests for each different tendon. It was found proportionality within regions. The initial strain, the toe region and the yield point corresponded to 25, 50 and 80% of the strain failure respectively Figure 2. That strain ratios were observed in all tendons excepting PB that presented a large yield region.

Figure 2: Typical stress-strain curve for tendon tissue.
Tibialis and peroneus tendons showed strain values twice than flexor and extensor toes tendons for each region. The linear region in flexor-extensor tendons were from strains values of 5.2% to 8.6% and in inversion-eversion tendons from strains values of 10.8% to 16.7%.

3.3 Young’s modulus

The summary of all trial curves are depicted in Figure 3. The points represent the failure point of the stress–strain curve of each tendon tested. Therefore, they are connected to the origin with a characteristic tendon curve.

Two different behaviors can be distinguished. On the one hand, flexor-extensor tendons (EDB, EDL, FDB, FDL, EHL, FHL) characterized by lower strain failure and higher Young’s modulus. The average curve of those tendons is graphed with a black solid line. On the other hand, inversion-eversion tendons (TA, TP, PB, PL) presented larger strains limits and lower Young’s modulus. The average stress-strain curve of those tendons is graphed with a red solid line. The flexor-extensor tendons presented an average Young’s modulus of 382MPa against 174MPa for inversion-eversion tendons. Those different behaviors were also observed in the strain values for each region of the stress-strain curve showed in Table 2.

There was no significant difference in any parameter analyzed between tendon toes and between flexors and extensors. Neither intersubject variability was found. At macroscopic failure, the fiber breakage could be heard.

![Figure 3: Failure points of all trials, sorted by tendon activity.](image-url)
4 DISCUSSION

In the present study, the tensile response of the human foot tendons was obtained in vitro. Young’s modulus is the input parameter for linear models in computational simulation which relates the stress to the strain. Therefore, significant differences in Young’s modulus denote distinct mechanical behavior. From uniaxial mechanical tests a significant difference ($p<0.001$) were found. Flexor-extensor toes tendons have doubled Young’s modulus than inversion-eversion tendons. In practical terms, this means that tibialis and peroneus tendons are more stretched for the same tensile stress.

The main experimental challenges found during testing were the measurement of the smallest CSA’s and the clamping of the samples to the testing machine. For measurement of tendon CSA, caliper was chosen because of its ease of use and its sufficient level of accuracy and repeatability. The critical point was measure the thinner tendons because of the sensibility of the measurement where little variations in the CSA have high impact in the calculus of the material properties. In fact, tendons with CSA smaller than 1.5mm$^2$ such as Lumbricals tendons were not included in the study for this reason. Blanton and Biggs [11] used a planimeter over a thirty magnifications projected area to determine tendon CSA. The CSA we measured were similar to those reported in that study, Figure 4. For avoiding slipping of the tendon samples, especially the thicker specimens, screw lock clamps were specifically designed to perform the tests. The inner sides of the clamps were milled with small holes to improve the grip. Due to the impact of the clamping system stress concentrations were expected in the samples. Therefore, in physiological position, values of tendon failure may be under-estimated.

Blanton and Biggs [11] also reported ultimate tensile stresses of lower flexor and extensor tendons separately. They found a similar range for both groups between 9-55MPa with mean stress failure of 26.2MPa and 31.2MPa for extensor and flexor tendons respectively compared with our result in a range of 7-77MPa with mean values of 22.8MPa for extensor tendons and 26.1MPa for flexor tendons.

![Figure 4: Comparison of CSA measurements reported by Blanton and Biggs and the present study.](image-url)
Other previous study measured mechanical properties of EDL in vitro [9]. The stress and strain failures reported in that study were much higher than our values. Schechtman & Bader [9] did not report Young’s modulus, but it can be inferred from their typical stress-strain curve, mean ultimate tensile stresses and mean strain failure results that they obtained values that doubled our measurements for that particular tendon. The differences of results are likely account of test speed. Schechtman & Bader [9] performed quasi-static tests at strain rates of 1% s$^{-1}$ while the tendons in the present study were tested about 10 to 20 times faster.

The material properties measured in the present study are in the range of values reported for other tendons. The results obtained can be utilized in different practical settings: In foot computational simulations to proper calculation of the coefficients for tendon constitutive models, in estimating overuses injuries during athletic activities, in designing of synthetic materials for tendon replacement and in reconstruction of tendons such as reparation ruptured Achilles tendon.

5 CONCLUSIONS

Two major finding were obtained from the experiments.

- Flexor-extensor toes tendons and inversion-eversion tendons present different mechanical properties. The former have higher Young’s modulus and ultimate tensile stresses and latter exhibit larger strain failures.
- Within stress-strain curve proportionality between regions was found. The initial strain, the toe region and the yield point corresponded to 25, 50 and 80% of the strain failure respectively.

6 LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSA</td>
<td>Cross-Sectional Area</td>
</tr>
<tr>
<td>EDB</td>
<td>Extensor Digitorum Brevis</td>
</tr>
<tr>
<td>EDL</td>
<td>Extensor Digitorum Longus</td>
</tr>
<tr>
<td>EHL</td>
<td>Extensor Hallucis Longus</td>
</tr>
<tr>
<td>FDB</td>
<td>Flexor Digitorum Brevis</td>
</tr>
<tr>
<td>FDL</td>
<td>Flexor Digitorum Longus</td>
</tr>
<tr>
<td>FHL</td>
<td>Flexor Hallucis Longus</td>
</tr>
<tr>
<td>PB</td>
<td>Peroneus Brevis</td>
</tr>
<tr>
<td>PL</td>
<td>Peroneus Longus</td>
</tr>
<tr>
<td>TA</td>
<td>Tibialis Anterior</td>
</tr>
<tr>
<td>TP</td>
<td>Tibiales Posterior</td>
</tr>
</tbody>
</table>

7 ACKNOWLEDGE

The authors gratefully acknowledge the support of the Ministry of Economy and competitiveness of the Government of Spain through the project DPI2013-44987-R.
8 BIBLIOGRAPHY


USING 1D-IRBFN METHOD FOR SOLVING A HIGH-ORDER NONLINEAR DIFFERENTIAL EQUATION ARISING IN MODELS OF ACTIVE-DISSIPATIVE SYSTEMS

DMITRY V. STRUNIN\textsuperscript{1}, DUC NGO-CONG\textsuperscript{2} AND RAJEEV BHANOT\textsuperscript{3}

\textsuperscript{1}Computational Engineering and Science Research Centre
and Faculty of Health, Engineering and Sciences
University of Southern Queensland
Toowoomba, QLD 4350, Australia
strunin@usq.edu.au

\textsuperscript{2}Computational Engineering and Science Research Centre
University of Southern Queensland
Toowoomba, QLD 4350, Australia
duc.ngo@usq.edu.au

\textsuperscript{3}Computational Engineering and Science Research Centre
and Faculty of Health, Engineering and Sciences
University of Southern Queensland
Toowoomba, QLD 4350, Australia

Key words: 1D-IRBFN Numerical Method, Nonlinear Active-Dissipative PDE

Abstract. We analyse a nonlinear partial differential equation modelling reaction-diffusion systems with nonlocal coupling and reaction fronts of gasless combustion. The equation is of active-dissipative type, nonlinear, with 6th-order spatial derivative. To numerically solve the equations we use the one-dimensional integrated radial basis function network (1D-IRBFN) method. The method has been previously developed and successfully applied to several problems such as structural analysis, viscous and viscoelastic flows and fluid-structure interaction. A commonly used approach is to differentiate a function of interest to obtain approximate derivatives. However, this leads to a reduction in convergence rate for derivatives and this reduction is an increasing function of derivative order. Accordingly, differentiation magnifies errors. To avoid this problem and recognising that integration is a smoothing process, the proposed 1D-IRBFN method uses the integral formulation, where spectral approximants are utilised to represent highest-order derivatives under consideration and then integrated analytically to yield approximate expressions for lower-order derivatives and the function itself. Our preliminary results demonstrate good performance of the 1D-IRBFN algorithm for the equation under consideration. Numerical solutions representing travelling waves are obtained, in agreement with the earlier studies.
1 INTRODUCTION

The following model presents a single-equation simulation tool for certain type of active systems with dissipation,

$$\partial_t u = -A(\partial_x u)^2 \partial_x^2 u + B(\partial_x u)^4 + C\partial_x^6 u,$$

(1)

A, B, C > 0. In particular, Eq. (1) simulates combustion waves (fronts) having the shape of one or more steps [1] and developing instabilities in nonlocal reaction-diffusion systems [2]. In the context of combustion waves, u stands for the distance, measured along, say, axis z, passed by the combustion front through a hollow cylinder, as a function of the coordinate x and time t. The equation generates a rich variety of dynamical regimes, the most spectacular of which is the spinning wave illustrated by Fig. 1(a),(b). The graph (b) shows the wave solution of (1) simulating the experiment (a). On the graph the cylinder is rolled out into a plane and two periods are shown. The moving x locations of the steepest sections in figure (b) correspond to luminous spots spinning along the cylinder, due to extremely high temperatures.

![Figure 1](image)

**Figure 1:** (a) A post-combustion trace on a hollow cylinder. (b) A running spinning wave solution of Eq. (1) (five successive shapes) evolved from a random initial condition [1].

Previously Eq. (1) has been solved numerically in 1D – with one independent spatial dimension x – using the spectral Galerkin method [1] and in 2D using a straightforward finite difference scheme. In this paper we aim to apply a different numerical method, which has proved to be fast and accurate in a number of problems, namely the 1D-IRBFN method. First we would like to test the method. Eq. (1) is unlikely to allow analytical solutions because of the nonlinearity, so we intend to obtain a single-step travelling wave and compare it with the results shown in Fig. 1. In this figure we see a train of fronts, each comprising a sharp step followed by a long nearly flat shoulder. Although the shoulders appear inclined, their slopes...
are much flatter than the abrupt steps. The speed of the front or steps, their height and width are controlled by the equation and not initial conditions. This is a consequence of the steps being formed as a result of the balance between the energy release in the system, represented by the term \(-A(\partial_xu)^2\partial_x^2u\), and the dissipation, represented by the term \(C\partial_x^6u\). The term \(B(\partial_xu)^4\) carries the function of the bridge between the above two. Our modelling in this paper is the first step in a new series of numerical exercises targeting various dynamical regimes generated by Eq. (1).

Note that by re-scaling \(t\), \(x\) and \(u\), Eq. (1) can always be transformed into a canonical form where all the coefficients, \(A\), \(B\), and \(C\), become equal to 1.

2 THE NUMERICAL METHOD

The 1D-IRBF and IRBF-based methods have been successfully verified previously through several engineering problems such as turbulent flows [3], laminar viscous flows [4, 5, 6], structural analysis [7], and fluid-structure interaction [8].

Radial basis function networks (RBFNs) have been known as powerful high-order approximation tools for scattered data [9]. A function \(f(x)\), to be approximated, can be represented by an RBFN as

\[
f(x) \approx u(x) = \sum_{i=1}^{N} w_i G_i(x),
\]

where \(x\) is the input vector, \(N\) the number of RBFs, \(\{w_i\}_{i=1}^{N}\) the set of network weights to be found, and \(\{G_i(x)\}_{i=1}^{N}\) the set of RBFs. According to Micchelli’s theorem, there is a large class of RBFs, e.g. the multiquadric, inverse multiquadric and Gaussian functions, whose design/interpolation matrices obtained from (2) are always invertible. It has been proved that RBFNs are capable of representing any continuous function to a prescribed degree of accuracy. Furthermore, according to the Cover theorem, the higher the number of RBFs used, the more accurate the approximation will be, indicating the property of “mesh convergence” of RBFNs. Among RBFs, the multiquadric functions \((G_i(x) = \sqrt{(x - c_i)^T(x - c_i) + a_i^2})\), \(c_i\) - the centre and \(a_i\) - the width) are ranked as the most accurate and possess an exponential convergence with the spatial discretisation refinement.

The application of RBFNs for solving partial differential equations has received wide attention over the past decades (e.g. [10] and references therein). The usual approach [11, 12] is to differentiate (2) as often as required to obtain approximate derivatives of \(f(x)\). If the error in \(f(x)\) is \(O(h^s)\), where \(h\) is the mesh size and \(s > 0\), the error in the \(n\)th derivative of \(f(x)\) is \(O(h^{s-n})\). In other words, there is a reduction in convergence rate for derivatives and this reduction is an increasing function of derivative order. Thus, differentiation will magnify any error that might exist in the approximation of \(f(x)\).

To avoid this problem, recognising that integration is a smoothing process, the integral formulation was proposed [13, 14], where spectral approximants (e.g. RBFNs) are utilised to represent highest-order derivatives under consideration and then integrated analytically to yield approximate expressions for lower-order derivatives and the function itself, to construct the ap-
proximations for the field variables in a problem. This approach was called integrated radial basis function networks or IRBFN. Although RBF methods can be easily implemented in a truly meshless manner based on scattered data points, it proves very efficient and effective to discretise a domain using Cartesian grids. Thus, the purpose of using integration (a smoothing operator) to construct the approximants is to avoid the reduction in convergence rate caused by differentiation, and also to improve the numerical stability of a discrete solution.

The integration process naturally gives rise to arbitrary constants that serve as additional expansion coefficients, and therefore facilitate the employment of some extra equations in the process of converting the RBF weights into the function values as illustrated above. This distinguishing feature of the integral formulation provides effective ways to overcome well-known difficulties associated with conventional differential approaches: (i) the implementation of multiple boundary conditions [15]; (ii) the description of non-rectangular boundaries on a Cartesian grid [16]; (iii) the imposition of high-order continuity of the approximate solution across subdomain interfaces [17]; and (iv) the incorporation of nodal derivative values into the approximations via compact IRBFN stencils (C-IRBFN) [18, 19]. The ability of the IRBFN methods to capture very sharp gradients, which is highly desirable for Eq. (1), has been demonstrated with the effective simulation of shockwave-like behaviours as in the dynamic strain localisation in a quasi-brittle material subjected to a sudden step loading [20].

3 RESULTS OF THE NUMERICAL EXPERIMENT: SETTLING OF THE TRAVELLING FRONT

Figures 2–3 show a sequence of snapshots displaying the solution of Eq. (1) evolving from the initial condition chosen in the step-like (front-like) form

\[ u(x, 0) = 8 \exp\left[-(x - 5)^2\right] \text{ for } x \geq 5, \]

\[ u(x, 0) \equiv 8 \text{ for } x < 5. \]

The idea is to help the solution curve acquire the step-like shape hinted by Fig. 1, although we realise that ultimate settled configuration of the solution will not depend of the initial condition.

In our numerical experiments we chose not to transform the equation to the canonical form, in order to be able to adjust the coefficient values as necessary, for example to increase or reduce the energy release in the system in hope to achieve a self-sustained balance between the release and dissipation. The parameter values were: \( A = 5, B = 1, C = 1 \), the number of nodes 401, the length of the \( x \)-interval 10, time step \( 2 \cdot 10^{-6} \). The \( x \)-interval was constantly shifted to the right to follow the main kink, which, according to the initial shape, was to move to the right. When looking at Fig. 1(b), if one stretches the \( x \)-axis enough, the long shoulders would become nearly horizontal. In fact each shoulder could stretch to infinity on both sides of an isolated single step. Aiming at such a solution, we adopt the boundary conditions

\[ \partial_x u = 0, \quad \partial_x^2 u = 0, \quad \partial_x^3 u = 0 \]

on the left and right ends of the (moving) \( x \)-interval.
Figure 2: Evolving solution of Eq. (1) via the 1D-IRBFN method.
Figure 3: Evolving solution of Eq. (1) via the 1D-IRBFN method.
The experiment showed that after some period of transitional evolution lasted from \( t_0 = 0 \) to about \( t = 0.03 \), the curve practically ceased changing in shape. Continuing the experiments further gave the same frozen shape of the solution moving with constant speed to the right. Looking at the solution displayed in Fig. 1(b) we see the same characteristic tale of ripples in front of the main kink. They were expected to form, caused by the high order of the dissipation acting in the system. Immediately in front of the main kink sits a shorter one, followed, as we look from left to right, by barely distinguishable smaller ripples. The height of the main kink relative to its neighbour is about 4:1 or a bit higher, for the both figures. The fact that the solution settles into a steady moving shape of correct proportions correlates with the earlier result.

4 CONCLUSIONS

We applied the 1D-IRBF numerical method to solve Eq. (1) simulating spinning combustion fronts and oscillations in certain class of reaction-diffusion systems. To our satisfaction, the method produced a similar shape of the settled travelling front as the earlier study [1]. We plan to use the 1D-IRBF approach to study more complicated regimes such as co-directed motion of several fronts, collision of counter-directed fronts etc.

REFERENCES


NUMERICAL SIMULATION OF COOLING AND FREEZING PROCESSES FOR THE CRYOPRESERVATION OF BIOLOGICAL MATERIAL IN LIQUID NITROGEN

VICTORIA SANTOS1,3, MARINA SANSINENA2,3, JORGE CHIRIFE2 AND NOEMÍ ZARITZKY1,3

1 Depto. de Ingeniería Química, Facultad de Ingeniería, Universidad Nacional de La Plata and Centro de Investigación y Desarrollo en Criotecnología de Alimentos www.cidca.org.ar (CONICET-UNLP) Calle 47 y 116, La Plata 1900, Argentina. mvsantosd@gmail.com


Key words: boiling regimes, nucleate boiling, film boiling, cryopreservation, surface heat transfer coefficients, liquid nitrogen

Abstract. The knowledge of the cooling and/or freezing rates in a biological sample during cryopreservation procedures is of major importance. Specifically, the measurement of temperature changes in the sample can be used to determine whether a sample is vitrified or undergoes phase transition. When dealing with liquid nitrogen (LN), having a saturation temperature at atmospheric pressure, $T_{\text{sat}} = -196 \, ^\circ\text{C}$, the temperature difference between the fluid and the sample, is large enough to cause boiling of LN, entering into the film boiling regime. This determines a heat flux from the object to LN causing the latter to boil in the immediate vicinity of the object and creating a pocket of nitrogen vapor around the object which acts as an “insulator” and retards further heat transfer. Film boiling is also referred to as the “Leidenfrost effect”. Boiling curves for a specific cryobiological system are scarcely found in the literature due to the small dimensions of the devices used in the process and the experimental limitations. The boiling curve which corresponds to variation of the heat flux ($q$) as a function of the excess wall temperature ($\Delta T = T_{\text{wall}} - T_{\text{sat}}$) depends on several relevant factors such as: i) the type of cryogenic fluid used and its thermal properties, ii) the material in contact with the fluid and its roughness which affects the nucleation sites and the Leidenfrost Temperature, iii) the relative position of the solid in the cryogenic fluid (vertical, horizontal, or in angle), and iv) the geometry of the solid (plates, sphere, cylinders, or irregular shapes). In this work the different heat transfer coefficients of straws filled with ice were experimentally determined by measuring time-temperature curves when they were immersed in LN; this allowed to observe the existence of different boiling regimes. The application of a numerical finite element program using the software COMSOL was used to predict the time-temperature curves and to obtain the surface heat transfer coefficients corresponding to each boiling regime. Independent experiments were carried out using straws that contained a biological fluid (semen+extender) which undergoes a phase change transition
(freezing), to further validate the surface heat transfer coefficients for film and nucleate pool boiling. The program takes into account the variable thermo-physical properties of the biological sample; this constitutes a highly non-linear mathematical problem. Additionally the numerical program allowed the determination of the range of $\Delta T$ where each regime develops, and the Leidenfrost temperature where the transition of film into nucleate pool boiling regime occurs. A good agreement was obtained between the experimental temperature profiles and the numerical predictions during cooling of straws containing both, ice and bovine semen + extender, confirming the reliability of the numerical results. Moreover, these results were corroborated with literature correlations in terms of dimensionless numbers. The numerical program is an important tool in order to correctly assess the heat transfer process and optimize the cryopreservation of straws filled with biological fluids.

1 INTRODUCTION

Liquid nitrogen is frequently used for the cryopreservation of biological samples. The measurement of temperature changes as the cooling process evolves can be used to calculate cooling rates and to determine whether a sample is vitrified or undergoes phase transition. When dealing with liquid nitrogen, having a saturation temperature at atmospheric pressure, $T_{sat}=-196 \, ^\circ C$, the temperature difference between the fluid and the sample, is large enough to cause boiling of the liquid entering into the film boiling regime \[1, 2\]. This determines a heat flux ($q$) from the object to liquid nitrogen creating a pocket of nitrogen vapor around the solid which acts as an “insulator” retarding further heat transfer. Film boiling is also referred to as the “Leidenfrost effect”. The object will cool down, rather slowly due to the low heat transfer rates and the “minimum heat flux” point will be reached at the Leidenfrost temperature ($T_L$). Vapor film will then break off while the heat flux progressively increases as transition to the nucleate boiling regime is established. This event is characterized by a steep increase of the heat flux up to a point called the “maximum heat flux” \[1, 3\].

The boiling curve which corresponds to variation of the heat flux ($q$) as a function of the excess wall temperature ($\Delta T=T_{wall} - T_{sat}$) depends on several relevant factors such as: i) the type of cryogenic fluid used and its thermal properties, ii) the material in contact with the fluid and its roughness which affects the nucleation sites and the $T_L$, iii) the relative position of the solid in the cryogenic fluid (vertical, horizontal, or in angle), and iv) the geometry of the solid (plates, sphere, cylinders, or irregular shapes).

There is a lack of experimental information concerning boiling curves in devices such as plastic straws containing biological samples. Plastic French straws are widely used in cryobiology and immersion in liquid nitrogen is a common procedure in cryopreservation of biological samples. Specifically, there exists no certain information whether the cooling process is governed entirely by a film or by a nucleate boiling regime, or a combination of both processes. Additionally it is important to assess the range of $\Delta T$ where each regime develops. Since nucleate boiling regime is characterized by high surface heat transfer coefficients it would be advantageous in order to reach vitrification or higher cooling rates for the process to take place during this regime.

The numerical modeling of the straws when plunged in liquid nitrogen enables the prediction of the surface heat transfer coefficients that characterize each boiling regime, as well as the temperature range where film and nucleate pool boiling develops. This
information is crucial to the understanding of the phenomena since the Leidenfrost
temperature (that is the point where a transition from film into nucleate boiling is produced),
depends on several factors. Additionally, a numerical model is useful because experimental
measurements are very difficult due to the small dimension of the devices.

The objectives of the present study were: i) to experimentally determine the time-
temperature curve of an ice-filled French straw when immersed in liquid nitrogen and to
define the boiling regimes that govern the process by applying a numerical finite element
program that calculates the variable surface heat transfer coefficients; ii) to validate the h
values using a French straw filled with a model biological fluid system (bovine semen-
extender) which undergoes freezing upon cooling with thermo-physical properties that are
temperature dependent and iii) to corroborate the h values with literature correlations based on
the Nusselt number for similar systems.

2 MATERIALS AND METHODS

2.1 Experimental Measurements

Cylindrical plastic straws used for bovine semen cryopreservation were obtained from AB
Technology, Inc., Pullman, Washington, USA). The average external diameters, length and
thickness were D=2.805mm, L=124mm, e=0.21mm, respectively.

In order to determine the surface heat transfer coefficients at the straw-LN interface, the
straw was filled with ultra-pure, reverse-osmosis filtered water (Milli-Q, Milipore
Corporation, MA, USA) that freezes at 0 ±0.3ºC. The time-temperature curve was recorded
using a thermocouple type T (Copper-Constantan) inserted in the central axis of the straw,
containing initially pure water. The straw containing pure water was slowly swirled in
Nitrogen vapor over liquid nitrogen in a Dewar tank to generate the ice crystals. The ice has
well-known thermophysical properties in the cooling range of -2 to -194ºC [4, 5]. Once the
freezing plateau region was completed the straw was maintained in a thermostatic bath using
a solution of ethylene glycol-water (30%v/v) at a temperature below -2ºC and above -9 ºC.
After equilibration the straw was rapidly plunged in liquid nitrogen using the Dewar canisters.
The advantage of using ice in the straw is to avoid phase changes and in this way the ice
sample contained in the straw experienced only cooling during the temperature range of -3 ºC
to -194ºC. The thermocouple was connected to an acquisition device (TESTO, Germany).
Figures a, b, and c) show the Dewar container with the TESTO used in the experiments and
how the thermocouple was coupled and mounted to the straw and the TESTO device.

2.2 Model system

The system (plastic straw and internal material) can be described as two concentric finite
cylinders of different substances: the inner material being either ice or biological cells (semen+
extender) and the plastic straw. The partial differential equations that represent the heat
transfer in the fluid that is submitted to the freezing process (Eq. 1) and the plastic support
(Eq. 2) considering radial and axial coordinates have been thoroughly described in Santos et
al. [6, 7] and are as follows:
\[
p(T) \frac{\partial T}{\partial t} + \left( k(T) r \frac{\partial T}{\partial r} \right) + \left( k(T) r \frac{\partial T}{\partial z} \right)
\]

(1)

\[
p_r \frac{\partial T}{\partial t} = \frac{\partial}{\partial r} \left( k_p r \frac{\partial T}{\partial r} \right) + \frac{\partial}{\partial z} \left( k_p r \frac{\partial T}{\partial z} \right)
\]

(2)

Figure 1a): Dewar filled with liquid nitrogen and PC connected to the acquisition device (TESTO) b) connection of TESTO to the plastic French straw and c) straw initially filled with pure water with the thermocouple threaded into the needle and mounted into the straw to avoid radial movement during the experiment.

The subscripts \( s \) corresponds to the ice or biological fluid and \( p \) to the plastic material. It can be noticed that the ice and biological suspension have thermal properties that are temperature dependent. In the case of the plastic support the thermo-physical properties \( (k_p, \rho_p, C_p) \) are considered constant. The modeling of the biological fluid constitutes a highly non-linear mathematical problem due to the abrupt change of the thermal properties with temperature during phase change transition [6, 7]. The initial temperature condition was considered uniform in both material domains.

The convective boundary condition at the interface plastic support-LN is:

\[-k_p \nabla T.n = h(T_{wall} - T_{ext})\]

(3)

where \( h \) is the surface heat transfer coefficient, \( k_p \) is the plastic thermal conductivity, \( T_{wall} \) is the variable surface wall temperature at the interface plastic support-liquid nitrogen, \( T_{ext} \) is the external temperature (in this case the saturation temperature of LN at atmospheric pressure), \( n \) is the normal outward vector, and \( \nabla T \) is the temperature gradient evaluated at the surface.

The effect of \( z \) axis in the cylindrical straw has been considered because the system has asymmetrical border conditions. During the manipulation of straws in Dewar flasks, the lower base surface is in contact with liquid nitrogen and at the top base (when the tank is not completely full) is in contact with nitrogen vapors which would mean a low heat transfer coefficient \( h \) at this interface (Fig.1). In order to include the worst situation possible we used...
a q=0 at the top base and convective flux at the lower base.

A subroutine that enables the introduction of a variable heat transfer coefficient h with time was coded in Matlab language in the main program which was originally generated in the commercial software COMSOL AB. This subroutine allows the prediction of an average h value for each boiling regime. Different heat transfer coefficients were introduced to simulate the temperature-time curve for the straw; then, experimental and predicted temperatures for each proposed h were compared. The heat transfer coefficient that minimized the Residual Sum of Squares given by Eq. (4) was selected.

\[ RSS = \sum (T_{exp} - T_{pred})^2 \]  

A triangular mesh using Lagrange elements of order 2 was applied to discretize the domains. The number of elements that constituted the mesh for the biological fluid and plastic support domains were 30656 and 23808, respectively (Fig. 2). All the numerical runs were tested for their computational speed; the maximum CPU time was less than 10 min using a PC Intel(R) Core(TM) i3 with a processor speed of 2.93 GHz and a RAM of 4GB.

The time discretization scheme used was a Backward Euler Differentiation (minimum order 1 and maximum order 5) with a tuning step having a maximum of 1 s and a minimum initial starting value of 0.0001s.

The numerical program calculates de temperature profile as a function of time, in the straw and the fluid during the freezing process, especially the temperature at the wall which is used in the prediction of the heat flux by applying the convective boundary condition.

![Figure 2: Spatial discretization of the biological fluid and plastic support domains using triangular elements. Boundary conditions applied, revolution body with a revolution axis at r=0 m.](image)

2.3 Thermophysical properties

In the case of ice-filled straws the thermo-physical properties used as input in the numerical program were as follows: the average specific heat was considered 1461.7 J/kgºC for temperatures below 0ºC. The thermal conductivity and density of ice as a function of
The thermo-physical properties of the semen+extender were calculated based on the composition of the biological fluid expressed in wet basis mass fraction: carbohydrates = 0.098, fat = 0.031, and protein = 0.027. The moisture content of the semen+extender was experimentally measured and found to be 84.4 %. Choi and Okos (1986) [4] equations were applied to estimate the thermal conductivity and density [8]. Specific heat and latent heat of ice melting of the semen + extender mixture were measured by using a Differential Scanning Calorimeter (DSC) (TA Instruments, New Castle, Delaware, USA) model Q100 controlled by a TA 5000 module. Experimental data of the apparent specific heat, where the sensible heat is merged with the latent heat, produced a curve with a large peak around the freezing point. However for mathematical modeling purposes the specific heat capacity function has to rise smoothly to a peak over a finite range of temperature [9]. Therefore based on experimental data a \( C_p \) vs. Temperature curve was constructed using a Gaussian and Heaviside function [9] considering the experimental values of the initial freezing point, the range of temperature change and the latent heat of melting. The equation used in the numerical program to represent the specific heat as a function of temperature is as follows:

\[
C_p(T) = C_{pff} + \frac{\Delta H_m}{T_s} f(H_{ea}) + D \Delta H_m
\]

(5)

where \( C_{pff} \) is the specific heat of the fully frozen state, \( T_s \) is the peak temperature point and \( D \) is a gaussian curve defined as

\[
D(T) = e^{-\frac{(T-T_s)^2}{4dT^2}}
\]

(6)

where \( dT \) is the half width of transition, that is the temperature difference from melt within which 84% of the latent heat occurs [9]; \( f(H_{ea}) \) is the Heaviside function which is a built in function in COMSOL Matlab environment that has continuous second order derivatives. This function enables the numerical finite element software (Comsol AB Multiphysics, 2005) to successfully deal with the abrupt change in the apparent specific heat of the sample with temperature. The specific heat of the biological fluid obtained by DSC and using Eq. 5-6 can be observed in Fig. 3.

Figure 3: Apparent specific heat of semen+extender sample experimentally obtained by DSC (- -) and using the Heaviside and Gaussian functions (▬).
3 RESULTS AND DISCUSSION

3.1 Surface Heat Transfer Coefficients for Ice filled Straws

The experiments with straws containing ice, that were plunged in liquid nitrogen, allowed the rigorous determination of the surface heat transfer coefficients avoiding any phase change in the cooling process (temperatures of ice ranging between -2 and -196°C). Fig. 4 a) shows the experimental temperatures versus the predicted values assuming two different hypotheses: i) considering film boiling regime during the entire cooling process (using a single and constant h value); ii) assuming film boiling for the first stage followed by nucleate boiling regime (higher h value); Fig. 4 b shows a replicate run. All temperature measurements using thermocouples have an experimental absolute error of 0.5°C.

![Figure 4](image)

**Figure 4:** a) Experimental time-temperature curves in the straw containing ice (●) at r= 0.0008 m, L =0.060 m. Predicted temperatures obtained from the computational simulation with variable h values ( - - ); b) results of a duplicate run. The curve (▬ ● ▬) shown in a) was obtained considering a single value of h=150 W/m² °C (film boiling regime) over the entire experiment.

As can be observed there is a lack of agreement between the experimental and predicted temperatures when a single constant value of h was used over the entire cooling curve (Fig. 4a). However good agreement was achieved considering two regimes during the cooling of straws. Note that since ice was used from the beginning, no plateau region corresponding to the latent heat of phase change is observed. The cooling curve showed an abrupt change in the slope which is attributed to the transition of film into nucleate boiling regime. This change in the cooling slope was also reported in other experiments with stainless steel blocks immersed in liquid nitrogen [10]. This change in the rate of cooling can be attributed to the presence of two different regimes, first film and then nucleate boiling. During pool film boiling the excess wall temperature (∆T) is at its maximum and the h that best fitted experimental results was h=150 W/m² °C. During nucleate boiling a rapid drop of the temperature was observed and the h value of 1300W/m² °C gave the best fit to the experimental temperatures.

3.2 Validation of the Surface Heat Transfer Coefficients using Straws filled with semen

Fig.5a shows experimental time-temperature values and the finite element numerical predicted curves for French straws filled with semen+extender assuming the same hypothesis as in Section 3.1: i) considering film boiling regime during the entire cooling process (using a
single constant h value); ii) assuming film boiling for the first stage (h =150 W/m² °C) followed by nucleate boiling regime (h = 1300W/m²°C).

As it can be noted, good agreement between experimental and predicted values was observed when assuming film boiling for the first stage (h=150 W/m²°C) and nucleate boiling regime (h=1300 W/m²°C) for the second stage. This observation reinforces the existence of two stages of boiling regime phenomenon. The curve shows a mild plateau region which can be attributed to the phase change transition of water into ice in the semen sample. Afterwards, an abrupt slope change in the cooling curve develops and a rapid drop in the temperature is observed which corresponds to the transition of film to nucleate pool boiling similarly as it was explained for the ice system. Fig. 5b shows another replicate obtained with French straws containing semen+extender fluid and the numerical predictions using the h values previously determined.

3.3 Literature correlations for film and nucleate pool boiling

The film boiling is characterized by a vapor film covering the solid surface frequently called a non-wetting situation [11]; due to the low thermal conductivity of the N2 vapor with respect to the liquid nitrogen there exists a very large temperature difference between the solid and the liquid nitrogen. The vapor insulates the straw from the liquid nitrogen and also prevents the nucleation of bubbles owing to the absence of solid/liquid contact limiting the heat flux transfer. The critical temperature transition (Leidenfrost temperature, T_L), may depend on the nature of the surface and specifically on its roughness [12]. Although film boiling is established when there is very large temperature difference between the surface and fluid, it is important to predict the minimum excess temperature difference that must exist to sustain that regimen.

During the transitional boiling near the minimum heat flux, a large amount of vapor is gluttied around the object and it tries to buoy outward. Near T_L the instability that occurs at the interface between vapor and liquid nitrogen leads to the formation of waves in the vapor.

Figure 5: a) Experimental time-temperature curves (●) in the straw containing the biological fluid that undergoes freezing at r= 0.0008 m; L =0.060 m. Predicted temperatures using the program with variable h values ( - - ); b) Results from a similar experiment at r=0.0004m L=0.0604m. The curve (▬ ● ▬) shown in a) was obtained considering a single value of h=150 W/m² °C (film boiling regime) over the entire experiment .
film that release the vapor into jets feeding overhead bubbles or slugs perpendicular to the surface separated by distances denominated Taylor critical wavelength \([13]\). This wavelength can be predicted as a result of balancing forces of surface tension against inertia and gravity \([14, 15]\) as:

\[
\lambda_c = 2\pi \left[ \frac{\sigma}{g (\rho_l - \rho_v)} \right]^{\frac{1}{2}}
\]  

(7)

In the correlations reported by Sakurai et al. \([16]\), Kida et al. \([17]\), Galloway and Mudawar\([14]\), for film boiling regime, the characteristic length is the critical Taylor wavelength \((\lambda_c)\) instead of the cylinder length. These waves were experimentally observed for tubes in liquid nitrogen during film boiling in Kida et al. \([17]\) and Hsu and Westwater \([18]\). The correlation proposed by Sakurai et al. \([16]\):

\[
Nu(\lambda_c) = 0.82 (M(\lambda_c))^\frac{1}{4}
\]

(8)

being Nu the Nusselt number defined as

\[
\frac{Nu}{k_u} = \frac{\lambda_c + k}{k_u}
\]

where:

\[
M(\lambda_c) = \frac{Gr_{\lambda,c} Pr_v L}{C_p(T_w - T_{sat}) (R Pr_L)^{1/2}}
\]

(9)

\[
Gr_{\lambda,c} = \frac{9}{4} \lambda_c^3 \frac{\rho_l - \rho_v}{\rho_v}
\]

(10)

where \(Gr\) is the Grashof number

\[
L' = L + 0.5 C p_v (T_{wall} - T_{sat})
\]

(11)

where \(L'\) is the latent heat of vaporization plus sensible heat of vapor phase

\[
Pr_v = \frac{C_p \rho_v}{k_v}
\]

(12)

\(Pr_v\) is the Prandtl number for the vapor phase

\[
Pr_L = \frac{C_p \rho_l}{k_l}
\]

(13)

\(Pr_L\) is the Prandtl number for the liquid phase

\[
S_p = \frac{(C p_v \Delta T)}{L' Pr_v}
\]

(14)

\[
E = \left( A + C \sqrt{B} \right)^{\frac{1}{3}} + \left( A - C \sqrt{B} \right)^{\frac{1}{3}}
\]

(15)

\[
A = \frac{1}{4} R^2 S_p^2 Pr_L^2
\]

(16)
It can be observed that in all the correlations the excess wall temperature \((\Delta T = T_{wall} - T_{sat})\) changes during the heat transfer process. The equations proposed for each mathematical correlation were coded in Matlab using the variable \(T_{wall}\) obtained from the numerical solution of the transient heat transfer equations as a function of time; values of \(h\) informed in the present work correspond to the average value over the time period assayed.

The correlation for vertical tubes applying Sakurai equations (8)-(19) estimated a characteristic length for liquid nitrogen \(L_c = 6.6\)mm, same value as was reported in Kida et al. [17] resulting in an average value of \(h = 148.2\) W/m² °C. From the obtained data it can be concluded that the Sakurai et al. [16] mathematical correlation was in accordance with the numerical fitted \(h\) value for the film boiling regime. In the present study the value predicted by the model was 150 W/m² °C, which is in excellent agreement with the previously calculated \(h\) using the literature correlations for vertical tubes.

The pool nucleate boiling of nitrogen has been studied more extensively and the values of \(h\) for the vertical tubes immersed in liquid nitrogen were obtained using the extensive data reported in Seader et al. [19]; in their work the \(h\) values were in the range of 946.3-7993 W/m² °C. In the present work the mean \(h\) value that fitted the experimental results for straws was 1300 W/m² °C that is within the range for the vertical cylinders immersed in liquid nitrogen under nucleate boiling regime.

3.4 Boiling Curve of Plastic French Straws when plunged in LN

The boiling curve is represented by plotting the heat flux \((q = h \Delta T)\) as a function of the excess wall temperature \((\Delta T = T_{wall} - T_{sat})\). This curve was determined using the predicted wall temperature of the straw \((T_{wall})\) and the \(h\) values (Fig. 6).

The heat flux for straws filled with ice (Fig. 6 a) have a \(q_{min}\) value of 19555 W/m² which corresponds to an excess wall temperature value of \(\Delta T = 130.4\)°C, being the wall temperature \(T_{wall} = -65.7\)°C. The stable film boiling regime extends up to the point where the heat flux is at its minimum \((q_{min})\) corresponding to \((\Delta T)_{q_{min}} = T_{wall} - T_{sat}\). This \((\Delta T)_{q_{min}}\) for the tested straw-liquid nitrogen system was in the range of 130-140°C, that corresponds to a critical wall temperature range (Leidenfrost temperature \(T_L\)) of -66 to -55°C.

In the case of the straw filled with the biological fluid (Fig. 6 b) the minimum flux, \(q_{min} = 21072\) W/m² was obtained at an excess wall temperature \(\Delta T = 140.5\)°C which corresponds to a wall temperature of -55.52 °C. When the straw surface temperature becomes lower than this value there is a transition from film boiling to nucleate boiling and the heat flux increases up to a maximum value, afterwards it decreases and \(\Delta T\) diminishes.

This \((\Delta T)_{q_{min}}\) depends on several important factors such as the type of cryogenic fluid used and its thermal properties, the material in contact with the fluid and its roughness which affect the nucleation sites and the transition or Leidenfrost temperature. Additionally other factors
are, the relative position of the immersed body in the cryogenic fluid (vertical, horizontal, or angle of surfaces in contact with the fluid), and the geometry of the solid (plates, sphere, cylinders, or irregular shapes).

4 CONCLUSIONS

Measurement of experimental time–temperature curves of plastic French straw filled with ice and immersed in liquid nitrogen was carried out enabling the estimation of the surface heat transfer coefficients by applying a numerical finite element program. A specific subroutine was coded in the program in order to calculate a variable $h$ with time since two boiling regimes, film followed by nucleate pool boiling, were detected to occur during the cooling process.

![Figure 6: Boiling curves for French straws filled with a) ice b) biological fluid (semen+extender) when immersed in LN.](image)

The existence of film and nucleate pool boiling regimes was validated by additional experiments using French straws filled with bovine semen + extender fluid, plunged in liquid nitrogen and comparing the time–temperature curves with the numerical predictions. Since this system freezes upon cooling, their temperature dependent thermophysical properties were taken into account in the computer code, as well as the variable surface heat transfer coefficients during the different boiling regimes. A good agreement was obtained between the experimental temperature profiles and the numerical predictions during cooling of straws containing bovine semen + extender, confirming the reliability of the previously determined heat transfer coefficients (150 W/m² K for film boiling and 1300 W/m² K for nucleate boiling). Calculated values of heat flux ($q$) vs. $\Delta T$, of straws placed vertically in liquid nitrogen were determined; this allowed the prediction of the Leidenfrost temperature, $q_{\text{min}}$, $q_{\text{max}}$ and the minimum $\Delta T$ necessary to sustain a film type behavior. The calculated surface heat transfer coefficients for film and nucleate pool boiling during immersion in liquid nitrogen were also corroborated using literature correlations for related systems and configurations.

From a biological point of view the measurement of the time-temperature curve is highly important and the numerical modeling allows to determine the effect of the straw material, mainly its thermal conductivity and the thickness on the cooling rate. Also if another
biological, cryoprotectant suspension + biological cells, is used the numerical model would be able to predict the cooling or freezing of this particular fluid when plunged in liquid nitrogen if the thermophysical properties were previously determined and used as input in the numerical model.

In future work the effect of surface roughness will also be studied. The potential increase of the nucleation sites by modifying the surface may enhance a faster transition to a nucleate boiling regime. This would help decrease the temperature range where the film boiling regime develops, therefore increasing the cooling rate of the sample.

The determination of the different boiling regimes that govern the cooling process when plunging straws in liquid nitrogen constitutes an important issue when trying to optimize cryopreservation procedures for biological systems. Furthermore, this information can lead to improvements in the design of cooling devices in the cryobiology field.

4 ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support from the following institutions that funded this research work: Facultad de Ciencias Agrarias, Universidad Católica Argentina, Ciudad Autónoma de Buenos Aires, Centro de Investigación y Desarrollo en Criotecnología de Alimentos (CIDCA-CONICET), Universidad Nacional de La Plata, and Agencia Nacional de Promoción Científica y Tecnológica (ANPCYT)-ARGENTINA. They also thank Mr. Vicente Valenti for his collaboration in the laboratory experiments.

REFERENCES


COMPLEX VARIABLE SENSITIVITY ANALYSIS OF THE RADIO FREQUENCY ABLATION PROCESS FOR CANCER TREATMENT

Juan F. Monsalvo and Manuel J. García

1Applied Mechanics Research Group
Universidad EAFIT, Medellín, Colombia
jmonsal3@eafit.edu.co (Juan F. Monsalvo)

Key words: Sensitivity analysis, Complex taylor series expansion (CTSE), Cancer treatment

Abstract. The complex Taylor series expansion (CTSE) method is applied in the sensitivity analysis of radio frequency ablation (RFA) procedures, in which the temperature distribution has to be accurately predicted in order to apply proper temperature values to tumor tissue and to avoid unwanted damage of healthy one. For this reason, the CTSE method was used to calculate local sensitivity. In this work, we solve a basic 2D model of the RFA process modelled by the bioheat transfer equation, and coupled with Joule heating equation. The accuracy, robustness and step-size independence are the main advantages of the CTSE method.

1 INTRODUCTION

The radiofrequency ablation (RFA) is a minimal invasive clinical method for the treatment of hepatocellular carcinoma that consist in the heating of biological tissues through the emission of an external electric field. The goal of the RFA treatment is to raise the temperature distribution of the tumor above a value of 43°C for a specific period of time where necropsy will be presented, while keeping the temperature of healthy tissues below this value in order to avoid unwanted damage.

The thermal modeling of the radiofrequency ablation treatment has been studied by several researchers, being Pennes [1], the first to introduce a mathematical equation explaining the heat transfer presented in a biological tissue, known as the bioheat transfer equation. Tunç et al. [2] applied the bioheat transfer equation in the analysis of hyperthermia treatments. Despite the great research efforts, until today the study and prediction of the temperature distribution is a big issue in the reliability and effectiveness of the treatment. Since the human body is a complex blend of tissue with heterogeneous...
Juan F. Monsalvo and Manuel J. García

electric and thermal characteristic. Hence, a thorough sensitivity analysis of each variable involved in the process could allow us to identify which ones impact the most on the tumor temperature distribution. The complex Taylor series expansion (CTSE) method was used to perform this analysis.

The complex Taylor series expansion method was suggested by Moler and Lynes [3] and Lynes [4]. Although, were Trapp and Squire [5] the firsts to obtain a method to approximate derivatives of real functions using complex variables. The CTSE method calculated the first derivative respect a parameter by perturbing the imaginary axes of the parameter. Therefore, there is not difference operations involved, avoiding the cancellation error present in the finite difference method, that allow the method to be step size independent, easy to implement and highly accurate. In this paper, a basic 2D model of the RFA process was modelled using the bioheat transfer equation coupled with Joule heating equation. The sensitivity was calculated using CTSE.

2 GOVERNING EQUATIONS

A simplified 2D model of the radiofrequency ablation is shown in Figure 1, where the tumor ($\Omega_2$) is bounded by healthy tissue ($\Omega_1$). The electric field will be generated from two electrodes located on the top and bottom of the healthy tissue, generating a voltage gradient. The top and bottom of the healthy tissue are covered with a cooling pad that conserve a constant temperature.

![Figure 1](image)

**Figure 1:** A simplified 2D model of a radiofrequency ablation treatment with an embedded tumor.

The RFA treatment can be expressed as the coupling of two physical phenomena. The first one is the Joule heating induced by the electric field generated from the electrodes, the second one is the bioheat transfer equation with an added source term, result of the Joule heating. The combination of this two models allow us to predict the temperature field distribution on all the domain. A quasistatic electric field approximation was used due that the wave length of the radiofrequency is much greater compared to the depth of the computational domain, therefore, the free charge continuity equation will take the
following form
\[ \nabla \cdot \vec{J} + \frac{\partial \rho}{\partial t} = 0 \] (1)

the electric current density \( \vec{J} \) \( [\text{A/m}^2] \) is related to the electric field through the Ohm’s law
\[ \nabla \cdot \sigma_e \vec{E} = 0 \] (2)

where \( \sigma_e \) \( [\text{s/m}] \) is the tissue electrical conductivity and \( \vec{E} \) \( [\text{V/m}] \) denotes the electric field. The electric field can be define as the gradient of the electric potential
\[ \vec{E} = \nabla \phi \] (3)

substituting Equation 3 into Equation 2 yields
\[ \sigma_e \Delta \phi = 0 \] (4)

From the Equation 4, it can be seen that the electric potential distribution inside the domain is described by the Laplace equation. A dirichlet boundary condition of an electric potential relative to ground is specified on each electrode. On the remaining external boundaries an ideal electrical isolation are assumed. The energy that is received into the tissue through the electric field is dissipated by the Joule heating.
\[ Q_j(x, y) = \sigma_e \vec{E} \cdot \vec{E} = \sigma_e |\nabla \phi|^2 \] (5)

The distribution of the temperature field inside the healthy and the tumor tissue is described by the bioheat transfer equation also know as the Pennes equation \([1]\). This equation has an added source term \( Q_j \) \( [\text{W/m}^3] \) product of the Joule heating. The steady version of the equation is used and can be defined as
\[ \lambda_e \nabla^2 T(x, y) + \rho_b C_b G_{be} [T_b - T(x, y)] + Q_{met} + Q_j(x, y) = 0 \] (6)
in which \( T \) \( [\circ \text{K}] \) denotes the temperature, \( \lambda_e \) \( [\text{W/m}^\circ \text{K}] \) is the thermal conductivity, \( \rho_b \) \( [\text{kg/m}^3] \) denotes the blood density, \( C_b \) \( [1/\text{kg}^\circ \text{K}] \) is the specific heat capacity of the blood, \( G_{be} \) \( [1/\text{s}] \) defines the perfusion rate, \( Q_{met} \) \( [\text{W/m}^3] \) denotes the metabolic heat source and \( Q_j \) \( [\text{W/m}^3] \) is the added heat source product of the radiofrequency treatment. The boundary conditions of the thermal model are as follow, a dirichlet condition is assumed constant along all the cooling pad and an adiabatic condition on the remaining external boundaries.
3 EVALUATION OF SENSITIVITY DERIVATIVES

A sensitivity analysis consists in measure the variation of the output of a system to any changes in the different parameters presented. The selection of the method for computing sensitivities isn’t an easy task and the key issues to take into account are the accuracy, the computational cost and in many cases it is also important that the differentiation scheme used can be easily implemented into the code. In this aspect the complex Taylor series expansion remain pretty easy to code, very accurate and doesn’t suffer from the subtraction error. The CTSE is based from the Taylor series expansion of \( f(x + i\Delta h) \) defined as

\[
f(x + i\Delta h) = f(x) + i\Delta h \dot{f}(x) - \Delta h^2 \frac{\ddot{f}(x)}{2} + \ldots
\]

(7)

taking the imaginary part of the Equation 7, neglecting the higher order terms and finally dividing it by the step size (\( \Delta h \)) yields

\[
\dot{f}(x) = \frac{Im(f(x + i\Delta h))}{\Delta h} + e(\Delta h)^2
\]

(8)

which is the first derivative with a second-order approximation. As can be seen the Equation 8 does not have a difference operation between two number, thus, avoid the subtraction error. Therefore, the step size can be extremely small in order to recover exact derivative.

3.1 Implementation

Although, the implementation of the CTSE method has on a finite element code is straightforward, some remarks have to be taken into consideration to accomplish this labor with the minimal effort. The conversion of the FEM solver into a complex number form involve the transformation of each number like parameters, coordinates and constants into their complex equivalent; and each complex number as \( a + ib \) can be expressed in a matrix form, where the real part is located on the diagonal of the matrix.

\[
a + ib = \begin{bmatrix} Re & -Im \\ Im & Re \end{bmatrix} = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}
\]

(9)

Therefore, the mathematics operations as the sum, the subtraction and multiplication between two complex numbers will be reduce at its corresponding matrix operations. Finally, the solutions of the different PDE will leave us with a vector as

\[
V_i = \begin{bmatrix} Re \\ Re \\ Re \\ Im \\ Im \\ Im \end{bmatrix}
\]

(10)
where the real part correspond to the normal output of the system, whereas, the imaginary part divided by the step size will give us the sensitivity.

4 NUMERICAL EXAMPLE

The geometrical domain used was based on the one used by Majchrzak [6] and is presented in Figure 1. Its dimensions are $0.08 \text{[m]}$ and $0.04 \text{[m]}$, the electrodes are located at $0.032 \text{[m]} \leq x \leq 0.048 \text{[m]}$ in the upper and lower part. The tumor is represented as a square domain which size is $0.015 \text{[m]}$ and its left inferior vertex has a coordinate of $x = 0.032 \text{[m]}, \ y = 0.016 \text{[m]}$. The values of the properties used for the healthy and the tumor tissue are presented in the following table.

<table>
<thead>
<tr>
<th>Tissue</th>
<th>$\lambda$ [W/m$^\circ$K]</th>
<th>$G_b$ [1/s]</th>
<th>$Q_{met}$ [W/m$^3$]</th>
<th>$\sigma$ [S/m]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Omega_1$</td>
<td>0.469</td>
<td>0.0005</td>
<td>4200</td>
<td>0.179</td>
</tr>
<tr>
<td>$\Omega_2$</td>
<td>0.6</td>
<td>0.002</td>
<td>42000</td>
<td>0.461</td>
</tr>
</tbody>
</table>

Source–Data from references [7], [8] and [9].

Table 1: Thermal and electrical properties of a healthy ($\Omega_1$) and a carcinoma ($\Omega_2$) human liver at a frequency of 100[KHz].

The value of the properties for the human blood are listed below.

<table>
<thead>
<tr>
<th>Blood</th>
<th>$T_b$ [°K]</th>
<th>$C_b$ [J/kg$^\circ$K]</th>
<th>$\rho_b$ [kg/m$^3$]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>310.15</td>
<td>4180</td>
<td>1000</td>
</tr>
</tbody>
</table>

Source–Data from reference [10].

Table 2: Thermal properties of the human blood.

These properties are considered independent of the temperature. The boundary conditions for the electric potential are as follows

- $\phi_1 = 10 \text{[V]}$ on the upper electrode.
- $\phi_1 = -10 \text{[V]}$ on the lower electrode.
- Ideal isolation on the remaining external boundaries.

for the bioheat transfer the following boundary conditions were implemented

- $T_1 = 305.65 \text{°K}$ on the cooling pad.
- Adiabatic condition on the remaining external boundaries.
5 RESULTS

The solutions obtained for solving the partial differential equations that are involved in the radio frequency ablation process for the numerical example specified before, will be shown below. In Figure 2 the electrical potential field for the boundary conditions imposed can be appreciated, the figure is the result of solve the Equation 4. Using this result and on the basis of Equation 5 the heating source \( Q_e \) can be determined (see Figure 3).

Finally, solving the Equation 6 the temperature distribution is obtained (see Figure 4), the maximal temperature reached 42.9°C is below the therapeutic value recommended (43°C). Hence, some improvements in the temperature distribution can be done in order to reach this value in most part of the tumor region. Several options are presented, the most direct option would be to increase the difference in the electric potential of the two electrodes, but study the effects of the tissues properties and certain parameters have into the temperature field are more interesting. Furthermore, identify which property is the most critical in this kind of treatment could lead us to new kind of treatment, and a better understand of the RFA process.
5.1 Sensitivity analysis

The goal of the research is to determine the effect that the parameters and the tissues properties have on the temperature distribution. This information can be used to obtain a best treatment with a minimal collateral damage. So, a sensitivity analysis was done using CTSE. As the sensitivity found with the CTSE method for each property is an absolute sensitivity which cannot be compared; using a relative sensitivity is advised. Therefore, each sensitivity will be multiplied by its parameter and will be divided by the temperature as presented below.

\[ S_T^k = \frac{\partial T(x, y)}{\partial k} \frac{k}{T(x, y)} \tag{11} \]

The Equation 11 will give us a dimensionless value that can be used to compare the sensitivity obtained for each one of the properties analyzed. The relative sensitivity obtained for the thermal conductivity of the healthy (see Figure 5) and the carcinoma (see Figure 6) tissues.

**Figure 5:** Relative sensitivity with respect to the thermal conductivity of the healthy tissue $\lambda_1$.

It can be seen how the thermal conductivity of the carcinoma has a reduced effect that is bounded to the tumor region, and the thermal conductivity of the healthy tissue proceed like a heatsink, spreading the temperature around the domain.
6 CONCLUSIONS

The use of computational tools for cancer treatment have proved to be a promising new source of research and hope for a reliable and efficient way to improve the therapies in cancer treatment, despite the huge amount of work remaining to do the sensitivity analysis is reliable source for improving the distribution of the temperature field and achieve a successful treatment with the minimal collateral damage.

The complex Taylor series expansion has offered great results in the sensitivity analysis. It’s capabilities as an accurate, step size independent method and easily coding make it an outstanding option over the others traditional methods.

7 ACKNOWLEDGEMENT

The presentation to this congress was supported by the Universidad EAFIT.

REFERENCES


LATTICE-BOLTZMAN MODELING OF POROUS MEDIUM: APPLICATION TO MODELING FLOW DIVERTER STENTS

Gustavo A. Boroni¹, Javier Dottori¹, Alejandro Clausse¹, Ignacio Larrabide¹

¹Pladema - CONICET - UNICEN
Pinto 399, Tandil, Argentina
gboroni@exa.unicen.edu.ar, jdottori@exa.unicen.edu.ar, clausse@exa.unicen.edu.ar, larrabide@exa.unicen.edu.ar
www.pladema.net

Key words: Lattice-Bolzman, FVM, porous medium, flow diverter

Abstract. The treatment of Intracranial Aneurysms (IA) with Flow Diverter (FD) stents is becoming more common every day. Therefore, there is a need for understanding their behavior in a predictive way. This paper presents the implementation and the application of a Lattice-Boltzman porous media model previously described, for modeling cerebrovascular Flow Diverter devices stents. Numerical experiments where prepared using a Finite Volume software and used to adjust the numerical parameters of the LBM model. Preliminary results show good agreement between both models.

1 Motivation

Intracranial aneurysms (IA) are pathological dilatations of arteries inside the skull, which may rupture if not treated on time. The rupture of an IA follows by subarachnoid hemorrhage (SAH) with high morbidity and mortality rates [17]. The use of intravascular stents has become more frequent in such pathologies. Flow Diverters (FDs) are stents made of braided threads of a shape memory alloy, typically Nitinol. Their design is radically different from conventional stents. In particular, FDs are characterized by very thin wires (~40 μm), very small pores (~310 μm²) and a single or a multi-layer structure.

These devices are becoming more and more used in the treatment of IA [14, 10]. Still, there are many aspects of their behavior and effect on local hemodynamics that are yet not fully understood. The use of Computational Fluid Dynamics (CFD) can be of great benefit to understand the effect and implications of their use [12, 8, 9]. Due to its complex geometrical structure, modeling such devices imposes a number of limitations and requires a considerable computational effort for the simulation of relatively small fluid domains.
Recent studies reveal FD porosity is a major design parameter driving their ability to modify local aneurysmal flow [7]. Because of this, seems natural to model such devices using a porous medium model [3].

The present work describes the implementation and the application of a Lattice-Boltzman Method (LBM) porous medium model previously described [5] for modeling cerebro-vascular FD devices.

2 Materials & methods

In this work is described the application of an existing LBM porous medium model, which was adjusted using the measurements obtained form Direct Numerical Simulations (DNS), to model FD stents. The LBM porous medium model is adjusted to reproduce the conditions imposed by the FD on the flow. The LB model has been adjusted using a Direct Numerical Simulations (DNS) experiments. In the following, a brief introduction to the LBM and the corresponding porous model is provided.

2.1 Lattice-Boltzman Method

The LBM is a numerical method with great potential in CFD. Among its advantages the ease of direct parallelization, the flexibility to include any type of forces and internal interactions, and the capability to easily operate in complicated geometries are perhaps the most remarkable ones. LBM is particularly advantageous in simulation of flow in porous medium, multiphase and multicomponent flows, hemodynamics, among other domains [4, 2, 13].

LBM is an explicit method to solve transport equations. It is based on a collision (Eq. 1) and a streaming step (Eq. 2) based on discretizations of the Boltzmann equations, using a particles distribution density $f_i(\vec{x}, t)$, which evolves by

$$ f_i'(\vec{x}, t) = f_i(\vec{x}, t) + \frac{1}{\tau}[f_i^e(\vec{x}, t) - f_i(\vec{x}, t)] + \Delta t g_i(\vec{x}, t), \tag{1} $$

$$ f_i(\vec{x} + \vec{e}_i \Delta t, t + \Delta t) = f_i'(\vec{x}, t), \tag{2} $$

where $\Delta t$ is the time step, $\tau$ is a relaxation parameter that controls the viscosity, $g_i(\vec{x}, t)$ is an external source rate, $f_i^e(\vec{x}, t)$ is the called the equilibrium population or distribution, and the vectors $\vec{e}_i$ form a finite set of $Q$ lattice directions. In this work the D3Q19 model is used, which provides second order convergence to Navier-Stokes equations. A full description of the model can be found in [6].

The kinematic viscosity in units of $\Delta x^2/\Delta t$ is related to the relaxation parameter by

$$ v = \frac{1}{6} (2\tau - 1). \tag{3} $$

The source term $g_{\alpha}(x, t)$ is used here to introduce external forces $\vec{F}(\vec{x}, t)$ acting on each cell. It is designed as body force $\vec{F}(\vec{x}, t)$ can be applied in each cell, and the following
expression suggested by [11] will be used

\[ g_i = 3w_i \vec{F} \cdot \vec{e}_i. \]  

(4)

2.2 Modeling of porous medium with LBM

The porous medium was simulated by means of source terms accounting for the drag forces imposed by the FD mesh across the channel. The method adjusted here has been introduced elsewhere, please see [5] for further details and validation. The rationale behind this method is the use of Darcy’s law, which states that the volumetric drag force \( \vec{F}_D \) is given by

\[ \vec{F}_D = -\frac{\rho \nu}{\kappa} \overrightarrow{\vec{u}} \]  

(5)

where \( \kappa \) is the permeability of the medium.

It is convenient to introduce the permeability in LBM units to simplify numerical calculations. In this case of FD, the permeability is related to the porous medium geometry. The \( LB \) subscript will be used to indicate LBM units. In the following equations, vector notation and parameters are omitted for clarity. Thus, both systems are related as follows

\[ u = u_{LB} \frac{\Delta x}{\Delta t} \]  

(6)

\[ v = \frac{1}{3} (t - 0.5) \frac{\Delta x^2}{\Delta t} \]  

(7)

\[ \rho = \rho_{LB} \frac{m_{LB}}{\Delta x^3} \]  

(8)

where \( m_{LB} \) is the particle mass. If a force is imposed on each cell, then

\[ F = F_{LB} \frac{\Delta x}{\Delta t^2} \]  

(9)

The proposed model introduces a force in each porous medium cell. To obtain an expression of the permeability in LBM, let us consider the force equation

\[ F_{LB} = -K \ u_{LB} \]  

(10)

Then, for a cell with volume \( \Delta x^3 \), the equivalent force per unit volume is

\[ F_D = \frac{F}{\Delta x^3} \]  

(11)

Using Eqs. 5 to 11 can be obtained leads to the equation K parameter

\[ K = \frac{1}{\kappa_{LB}} \frac{(t - 0.5)}{3} \rho_{LB}. \]  

(12)
where $\kappa_{LB}$ is the permeability in LBM units, given by the natural unit conversion between systems

$$\kappa = \kappa_{LB} \Delta x^2.$$  

(13)

The simplest Darcy’s law implementation in LBM (Eq. 5) involves applying a uniform force in each cell of the porous region. Although this is a simple alternative, previous numerical tests showed poor agreement with DNS. We then propose a model that computes the Darcy force in designated reference points. A neighborhood is defined around reference points by a weighting function $\delta(\vec{x}, \vec{y})$. Each reference point introduces a source term $F_k(\vec{x}, t)$ to the cells located in its neighborhood, given by

$$F_k(\vec{x}, t) = -KN_k \delta(\vec{x}_k, \vec{x}) \vec{u}_k(\vec{x}, t)$$  

(14)

where $N_k$ is the number of LBM cells that constitute a unit cell of the permeable media, $\delta(\vec{x}, \vec{y})$ is the neighborhood weighting function and $\vec{u}_u$ is the weighted average reference point velocity according to

$$\vec{u}_k = \sum_x \vec{u}(\vec{x}) \ast \delta(\vec{x}, \vec{x}_k)$$  

(15)

The weighting function is a discretization of Dirac delta, similar to that used in immersed boundary methods. The force is distributed between neighboring cells located at a distance $r \Delta x$ in all directions from the reference point. An influence radius $r = 2$ was considered for a smooth application of the force on the fluid domain

$$\delta(\vec{x}, \vec{y}) = \delta_d(x_x, y_x) \delta_d(x_y, y_y) \delta_d(x_z, y_z)$$  

(16)

$$\delta_d(x, y) = \begin{cases} C(r) \left[1 + \cos \frac{z(x-y)}{r \Delta x}\right] & \text{if } \left|\frac{(x-y)}{r \Delta x}\right| < 1 \\ 0 & \text{otherwise} \end{cases}$$  

(17)

where $C(r)$ is a normalization factor and for $r = 2$ has a value of 1/4.

A Finite Volume CFD package was used to study the behavior of $\kappa$ in different conditions for physiological ranges of velocity and viscosity of IA blood flow. The experimental setup and numerical simulation environment used are explained in the next section.

2.3 Parameter identification - DNS experiments

Numerical experiments were performed to measure the pressure drop imposed by the FD mesh on the flow. These experiments were performed with the open-use Finite Volume (FV) code OpenFOAM (OpenFOAM, Bracknell, UK) [1] with the Immersed Boundary extension by Z. Tukovic and H. Jasak [15]. The mentioned code implements the Immersed Boundary method, which was used to model the FD mesh across a straight channels.

Although blood is a non-Newtonian fluid, for the Reynolds’ considered in aneurysmal blood flow simulation problem ($Re = 30 \sim 300$) its behavior can be safely assumed as Newtonian. Still, different dynamic viscosities where considered to account for the physiological range ($\mu = 0.001, 0.004$ and $0.015 \text{ Pa} \cdot \text{s}$).
The FV numerical experiments were performed on a channel with 20mm×20mm of cross section, with slip boundary condition on the channel lateral walls. A typical FD mesh, with 40µm threads crossing at 90 degrees separated by 310µm, was created and placed at the center of the channel and perpendicular to the flow direction. The hexahedral mesh generated with blockMesh (OpenFOAM, Bracknell, UK) mesh was refined at the location of the FD mesh to a size of 12.5 µm, to ensure that at least 3 elements covered the strut width.

The inlet boundary conditions were set to reproduce normal physiological conditions in a cerebral vessel of 0.04 mm² cross-section. Baseline flow was set with 0.1m/s uniform velocity at the inlet of the channel. Two additional BC models where ran for high (0.15m/s) and low (0.05m/s) inflow velocity, for a more complete coverage of the range of physiological velocities. The pressure was set to zero at the (single) outlet of the channel. This resulted in 9 simulations (3 for each boundary condition and 3 different viscosities).

3 Results

Simulation results were post-processed and hydro-static pressure was measured along the center of the channel before and after the position of the mesh. This measurement represents the pressure drop imposed by the FD mesh on the flow.
3.1 DNS results

Figure 1 shows simulation results for DNS ((a) and (b)). The velocity field is shown in (a) and the pressure field is shown in (b), both with the explicit FD mesh used in the DNS models. In (a), longitudinal plane is colored with velocity contours. Also, a cross sectional plane taken at the axis of the FD mesh is shown with a surface warped by the velocity field at each point and scaled by 0.001 (maximum velocities at that plane are in the order of 0.1 m/s). Figure b) shows the pressure field as volume rendering, where high pressure (rendered in red) can be observed upstream and low pressure (rendered as light transparent blue) appears downstream from the FD mesh.

Figure 3 shows the dimensionless results (empty squares) for the different experiments. The blue line corresponds to the relation between Re and pressure for $\mu = 0.015$ Pa·s, the red line for $\mu = 0.004$ Pa·s and the black line for $\mu = 0.001$ Pa·s. An exponential relation can be observed between the Reynolds’ number and $\Delta p$. An exponential is adjusted to this results and then used for the simulation of LBM porous medium models described above.

3.2 Parameter adjustment for the LBM porous medium

As observed from the experiments described in the previous section, the relation between the velocity and the pressure drop is non-linear for the velocity and viscosity ranges of interest. Then, a higher order approximation for $\kappa$ is required. The data obtained from the DNS experiments were adjusted using an exponential model of the form:

$$\Delta p = Au^B$$  \hspace{1cm} (18)

The values of $A$ and $B$ where adjusted, and the result was combined with Eq. 5. After computing expected $\Delta p$, the Darcy force per unit volume exerted by the FD is defined as

$$F_D = - \frac{\Delta p}{L}$$  \hspace{1cm} (19)

Finally, the effective permeability of the reference point can be computed by

$$k = - \frac{\rho \nu u}{F_D} = - \frac{\rho \nu u L}{\Delta p}$$  \hspace{1cm} (20)

This $\kappa$ value is computed in each reference point at each time step. For the studied viscosity values, the values for $A$ and $B$ are reported in Table 1. As expected, $B$ tends to 1 for greater viscosities and $A$ increases with viscosity.

3.3 LBM simulations set-up

As described above, a reference point is introduced in each LBM cell crossed by the FD. The first test was to reduce the size of computational domain. A grid resolution
Table 1: Values obtained for the exponential fit of parameters A and B

<table>
<thead>
<tr>
<th>$\mu$ [Pa·s]</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>0.4217</td>
<td>1.3955</td>
</tr>
<tr>
<td>0.004</td>
<td>0.5450</td>
<td>1.1456</td>
</tr>
<tr>
<td>0.015</td>
<td>1.2836</td>
<td>1.0339</td>
</tr>
</tbody>
</table>

Figure 2: Sample of different mesh sizes for the LBM cell grid (in number of cells in the cross section) compared the FD mesh.
analysis was performed varying the transversal grid size as a function of the number of rhombus shapes of the FD represented. In this case the grid sizes used in this analysis were $180 \times 180$, $90 \times 90$ and $45 \times 45$ corresponding to 16, 4, 1 FD mesh cells (rhombus), respectively (Figure 2). After computing the pressure gradient it was concluded that the effect of the grid resolution is negligible, and the profiles overlap. With this in mind, it was determined that using a grid size of $45 \times 45 \times 300$ would provide reasonable accuracy with a low computational cost.

The relaxation parameter was set to $\tau = 0.8$ and $\Delta x = 9.8052E-06$ (function of the grid size and the strut separation). Table 2 lists another parameter values used in the calculations.

### Table 2: Parameters used for the set-up of the LBM numerical simulations

<table>
<thead>
<tr>
<th>$\mu$ [Pa·s]</th>
<th>Inlet velocity [m/s]</th>
<th>$\Delta t$ [s]</th>
<th>Inlet velocity (LB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>0.05</td>
<td>1.02E-05</td>
<td>5.19E-02</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td></td>
<td>1.03E-01</td>
</tr>
<tr>
<td></td>
<td>0.15</td>
<td></td>
<td>1.56E-01</td>
</tr>
<tr>
<td>0.004</td>
<td>0.05</td>
<td>2.55E-06</td>
<td>1.30E-02</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td></td>
<td>2.60E-02</td>
</tr>
<tr>
<td></td>
<td>0.15</td>
<td></td>
<td>3.89E-02</td>
</tr>
<tr>
<td>0.015</td>
<td>0.05</td>
<td>6.80E-07</td>
<td>3.46E-03</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td></td>
<td>6.93E-03</td>
</tr>
<tr>
<td></td>
<td>0.15</td>
<td></td>
<td>1.04E-02</td>
</tr>
</tbody>
</table>

### 3.4 Effects of Reynolds number on the pressure gradient

Figure 3 shows the dimensionless pressure gradient $\Delta p/(\rho u^2)$ as a function of the Reynolds number $Re$. It can be seen the typical velocity profile for laminar boundary layer. This is found only when the $Re$ are small [16]. These results are evaluated by comparing the FV and LBM methods, for cases with dynamic viscosity 0.001, 0.004 and 0.015 [Pa·s], while maintaining all other parameters constant. Table 3 lists the permeability values. In these figures it can be seen that the profiles have a slight difference, and this difference is about 0.05%.

### 4 Conclusions

In this work we developed a methodology to parametrize a LBM porous medium model using DNS simulation results. A series of DNS models representing a channel with an explicit FD mesh across the central section of the channel were where prepared and ran. From these simulations pressure gradients across the FD mesh where measured for
different inflow velocities and viscosities, all of which ranged in physiological values of cerebral blood flow velocity and viscosity. These values where used to set up and tune a porous medium model. The porous medium model was ran to verify the adjustment of the model. It was found that an exponential law can accurately reproduce the effect of the FD mesh. These experiments were set on the basis of a single FD mesh, with an angle of 90 degrees between threads. Further experiments for different FD meshes are ongoing work. Further analysis of the proposed methodology is needed to assess its behavior in other conditions and orientations of flow.

5 Acknowledgements

This research has been partly supported by the Universidad Nacional del Centro de la Provincia de Buenos Aires and CONICET, Argentina. J.D. has been funded by a PhD scholarship by CONICET. The support of these institutions is greatly appreciated.

REFERENCES


Table 3: Permeability values used in the LBM simulations

<table>
<thead>
<tr>
<th>$\mu$ [Pa $\cdot$ s]</th>
<th>Inlet velocity [m/s]</th>
<th>$m_{LB}$</th>
<th>$\Delta p$ [Pa]</th>
<th>$\kappa$</th>
<th>$1/\kappa_{LB}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>0.05</td>
<td>1.98E-15</td>
<td>6.80E-03</td>
<td>2.29E-03</td>
<td>4.19E-05</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td></td>
<td>1.66E-02</td>
<td>1.87E-06</td>
<td>5.11E-05</td>
</tr>
<tr>
<td></td>
<td>0.15</td>
<td></td>
<td>3.00E-02</td>
<td>1.56E-06</td>
<td>6.16E-05</td>
</tr>
<tr>
<td>0.004</td>
<td>0.05</td>
<td>3.19E-14</td>
<td>1.80E-02</td>
<td>3.46E-06</td>
<td>2.77E-05</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td></td>
<td>3.85E-02</td>
<td>3.24E-06</td>
<td>2.96E-05</td>
</tr>
<tr>
<td></td>
<td>0.15</td>
<td></td>
<td>6.22E-02</td>
<td>3.01E-06</td>
<td>3.19E-05</td>
</tr>
<tr>
<td>0.015</td>
<td>0.05</td>
<td>4.53E-13</td>
<td>5.70E-02</td>
<td>4.10E-06</td>
<td>2.34E-05</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td></td>
<td>1.20E-01</td>
<td>3.90E-06</td>
<td>2.46E-05</td>
</tr>
<tr>
<td></td>
<td>0.15</td>
<td></td>
<td>1.80E-01</td>
<td>3.90E-06</td>
<td>2.46E-05</td>
</tr>
</tbody>
</table>


EXISTENCE AND STABILITY OF MINIMIZERS OF MIXED ANISOTROPIC BV-\(L^2\) TIKHONOV-PHILLIPS FUNCTIONALS: APPLICATIONS TO IMAGE RESTORATION.

GISELA L. MAZZIERI\(^1,2\), RUBEN D. SPIES\(^1,3\) AND KARINA G. TEMPERINI\(^1,4\)

\(^1\)Instituto de Matemática Aplicada del Litoral, IMAL, CCT CONICET Santa Fe Colectora Ruta Nac. 168, Paraje El Pozo, 3000 Santa Fe, Argentina

\(^2\)Departamento de Matemática, Facultad de Bioquímica y Ciencias Biológicas, Universidad Nacional del Litoral, Ciudad Universitaria, Paraje El Pozo, 3000 Santa Fe, Argentina
glmazzieri@santafe-conicet.gov.ar

\(^3\)Departamento de Matemática, Facultad de Ingeniería Química, Universidad Nacional del Litoral, Santiago del Estero 2829, 3000 Santa Fe, Argentina
rspies@santafe-conicet.gov.ar

\(^4\)Departamento de Matemática, Facultad de Humanidades y Ciencias, Universidad Nacional del Litoral, Ciudad Universitaria, Paraje El Pozo, 3000 Santa Fe, Argentina
ktemperini@santafe-conicet.gov.ar

Key words: Inverse problem, ill-posedness, regularization, Tikhonov-Phillips, bounded variation, anisotropy.

Abstract. During the last two decades several generalizations of the traditional Tikhonov-Phillips regularization method for solving inverse ill-posed problems have been proposed. Many of these variants consist essentially in modifications of the penalizing term, which forces certain features in the obtained regularized solution (\([8], [13]\)). If it is known that the regularity of the exact solution is inhomogeneous it is often desirable the use of mixed, spatially adaptive methods (\([6], [9]\)). These methods are also highly suitable when the preservation of borders and edges is also an important issue, since they allow for the inclusion of anisotropic penalizers for border detection (\([15]\)).

In this work, we propose the use of a penalizer resulting from the convex spatially-adaptive combinations of classic penalizing \(L^2\) and anisotropic bounded variation seminorm. Results on existence and uniqueness of minimizers of the corresponding Tikhonov-Phillips functional are presented. Stability results of those minimizers with respect to perturbations in the data, in the regularization parameter and in the operator are also established. An application to image restoration problem is shown.
1 INTRODUCTION AND PRELIMINARIES

A linear inverse problem can be formulated in the form: find \( u \) such that
\[
Tu = v,
\]
where \( \mathcal{X} \) and \( \mathcal{Y} \) are two infinite dimensional normed spaces (usually Hilbert spaces of functions), \( T \) is a bounded linear operator with non-closed range between those two spaces, and \( v \) is the data, which is known or approximately known (with a certain error). Under these hypotheses it is well known that problem (1) is ill-posed, the Moore-Penrose pseudo-inverse of \( T \) is unbounded and therefore small errors in the data \( v \) may result in arbitrarily large errors in the approximations of \( u \) ([14]). For this reason, before attempting to solve problem (1), it must be “regularized”. That means essentially replacing the problem by a family or sequence of “well-posed” problems whose solutions converge (in an appropriate way) to a solution of the original problem (1). Undoubtedly, the most usual way of regularizing a linear ill-posed problem is by means of the Tikhonov-Phillips method, which can be formulated as a simple unconstrained minimization problem. In fact, the regularized solution obtained by the Tikhonov-Phillips method and a penalizer \( W \) with domain \( D \subset \mathcal{X} \), is the global minimizer over \( D \) (provided such a minimizer exists), of the functional
\[
J_{\alpha,W}(u) = \|Tu - v\|^2 + \alpha W(u),
\]
where \( \alpha > 0 \) is a constant called regularization parameter. The original method was proposed independently by Phillips and Tikhonov in 1962 and 1963 ([11] and [16]) using \( W(u) = \|u\|^2_\mathcal{X} \). Other penalizers can also be used to regularize the problem and in the last two decades, considerable research has been devoted studying what types of functionals can be used for that purpose and, given a problem, decide which one is more appropriate to preserve certain known properties of the exact solution.

For general penalizers \( W \), sufficient conditions guaranteeing existence, uniqueness and stability of the minimizers under different types of perturbations were found in [5, 8, 12, 13].

Some results of existence and uniqueness of minimizers of generalized Tikhonov-Phillips functional using the term of penalty
\[
W(u) = \alpha_1 \int_\Omega |\sqrt{1 - \theta(x) u(x)}|^2 \, dx + \alpha_2 \int_\Omega \|\theta(x)|A(x)\nabla u(x)|^2 \, dx
\]
where \( \alpha_1, \alpha_2 \) are positive constants, \( \theta(x) \) is a weighting function with values on the interval \([0,1]\) and \( A(x) \) is a matrix field, were shown in [10]. In this work we present some extensions of those results by allowing the weighting function \( \theta \) to take the extreme values 0 and 1 in sets of positive measure and also by considering different norms in the anisotropic penalizer in (3). Some of these results are anisotropic generalizations of similar results obtained in [9]. Note that the extreme case \( \theta(x) \equiv 0 \) corresponds to the classical Tikhonov-Phillips method, while for \( \theta(x) \equiv 1 \) one gets an anisotropic pure
BV method, with the classical bounded variation method corresponding to the case of $A(x) \equiv I$. Other choices of the matrix field $A$ are possible in order to induce anisotropy. The construction of this matrix field is a very important matter on which we shall not get any deeper here. We will only mention that there are several ways of constructing this so-called “anisotropic matrix field”, either from structural prior information or from the available data (see [4], [7]). The general case can then be thought of as a convex combination of a classical $L^2$ and an anisotropic BV penalizers.

In what follows $\Omega \subset \mathbb{R}^2$ is a bounded open convex set with Lipschitz boundary, $\mathcal{M}(\Omega)$ shall denote the set of all real valued measurable functions defined on $\Omega$ and $\hat{\mathcal{M}}(\Omega)$ the subset of $\mathcal{M}(\Omega)$ formed by those functions with values in $[0, 1]$.

**Definition 1.1** Given $\theta \in \hat{\mathcal{M}}(\Omega)$, a measurable matrix field $A : \Omega \to \mathbb{R}^{2 \times 2}$ and $p \in [1, \infty)$ we define the functional $W_{\theta, A, p}(u)$ with values on the extended reals by

$$W_{\theta, A, p}(u) = \sup_{\tilde{\nu} \in V_{\theta, A, p}} \int_{\Omega} -u \div (\theta A \tilde{\nu}) \, dx, \quad u \in \mathcal{M}(\Omega)$$

where $V_{\theta, A, p} = \{ \tilde{\nu} : \Omega \to \mathbb{R}^2 \text{ such that } (\theta A \tilde{\nu}) \in \mathcal{C}_0^1(\Omega) \text{ and } |\tilde{\nu}(x)|_{p^*} \leq 1 \forall x \in \Omega \}$ and $p^*$ is the conjugate dual of $p$.

**Definition 1.2** The functional $J_0$ (BV-seminorm) with values on the extended reals is defined by

$$J_0(u) = \sup_{\tilde{\nu} \in V} \int_{\Omega} -u \div \tilde{\nu} \, dx,$$

with $V = \{ \tilde{\nu} : \Omega \to \mathbb{R}^n \text{ such that } \tilde{\nu} \in \mathcal{C}_0^1(\Omega) \text{ and } |\tilde{\nu}(x)| \leq 1 \forall x \in \Omega \}$.

**Remark 1.3** Note that if $\theta(x) \equiv 1$, $A(x) \equiv I$ (the identity matrix) and $p = 1$ then $W_{\theta, A, p}(u) = J_0(u)$.

**Proposition 1.4** Let $W_{\theta, A, p}$ be as in (4). If $\theta, u \in W^{1,1}(\Omega)$ and $A : \Omega \to \mathbb{R}^{2 \times 2} \in W^{1,1}(\Omega, \mathbb{R}^{2 \times 2})$ symmetric, then

1. $W_{\theta, A, 1}(u) = \| \theta |A\nabla u|_1 \|_{L^1(\Omega)}$;
2. $W_{\theta, A, 2}(u) = \| \theta |A\nabla u|_2 \|_{L^1(\Omega)}$.

**Proof** We will only prove part 1. Part 2 can be proved following similar steps. To prove part 1, assume first that $\theta, u \in C^1(\Omega)$ and $A \in C^1(\Omega, \mathbb{R}^{2 \times 2})$ (by standard density arguments, the result can be proved for general $\theta, u \in W^{1,1}(\Omega)$ and $A \in W^{1,1}(\Omega, \mathbb{R}^{2 \times 2})$).
For all $\tilde{v} \in V_{\theta,A,1}$ it follows easily that

$$\int_{\Omega} -u \operatorname{div}(\theta A \tilde{v}) \, dx = \int_{\Omega} \nabla u \cdot \theta A \tilde{v} \, dx - \int_{\delta \Omega} (u \theta A \tilde{v} \cdot \vec{n}) \, dS$$

$$= \int_{\Omega} \nabla u \cdot \theta A \tilde{v} \, dx \quad \text{(since $\theta A \tilde{v}|_{\delta\Omega} = 0$)}$$

$$= \int_{\Omega} \theta A \nabla u \cdot \tilde{v} \, dx \quad \text{(since $A$ is symmetric)}$$

$$\leq \int_{\Omega} |\theta A \nabla u|_1 |\tilde{v}|_\infty \, dx$$

$$\leq \int_{\Omega} \theta A \nabla u \, dx \quad \text{(since $|\tilde{v}(x)|_\infty \leq 1 \forall x \in \Omega$)},$$

where $\vec{n}$ denotes the outward unit normal to $\delta\Omega$. Taking supremum over $\tilde{v} \in V_{\theta,A,1}$ it follows that

$$W_{\theta,A,1}(u) \leq \|\theta|A\nabla u|_1\|_{L^1(\Omega)}.$$

For the opposite inequality, define $\tilde{v}_* (x) = \operatorname{sign}(A \nabla u(x))$. Then one has that $|\tilde{v}_*(x)|_\infty \leq 1 \forall x \in \Omega$ and

$$\int_{\Omega} A \nabla u \cdot \theta \tilde{v}_* \, dx = \int_{\Omega} |\theta A \nabla u|_1 \, dx.$$

By convolving $\tilde{v}_*$ with a function $\varphi \in C_0^\infty(\Omega, \mathbb{R}^2)$ we can obtain a function $\tilde{v} \in C^1(\Omega, \mathbb{R}^2)$ such that $|\tilde{v}(x)|_\infty \leq 1 \forall x \in \Omega$ (and therefore $\tilde{v} \in V_{\theta,A,1}$ since $u$ and $\theta$ are in $C^1(\Omega)$ and $A \in C^1(\Omega, \mathbb{R}^{2 \times 2})$) for which the left hand side of (6) is arbitrarily close to $\int_{\Omega} A \nabla u \cdot \theta \tilde{v} \, dx$.

From this it follows immediately that

$$\int_{\Omega} -u \operatorname{div}(\theta A \tilde{v}) \, dx \geq \|\theta|A\nabla u|_1\|_{L^1(\Omega)}.$$

Then taking supremum over $\tilde{v} \in V_{\theta,A,1}$ we have that

$$W_{\theta,A,1}(u) \geq \|\theta|A\nabla u|_1\|_{L^1(\Omega)}.$$

Hence $W_{\theta,A,1}(u) = \|\theta|A\nabla u|_1\|_{L^1(\Omega)}$, as we wanted to prove.

The previous proposition induces us to formulate the following conjecture.

**Conjecture 1.5** If $\theta, u \in W^{1,1}(\Omega)$, $A : \Omega \to \mathbb{R}^{2 \times 2} \in W^{1,1}(\Omega, \mathbb{R}^{2 \times 2})$ symmetric and $p \in [1, 2]$, then $W_{\theta,A,p}(u) = \|\theta|A\nabla u|_p\|_{L^1(\Omega)}$.

## 2 MAIN RESULTS

In this section we will state our main results on existence, uniqueness and stability of minimizers of generalized Tikhonov-Phillips functionals with penalizers involving spatially varying combinations of the $L^2$-norm and of the functional $W_{\theta,A,p}$, under different
hypotheses on the function $\theta$ and the anisotropic matrix $A(x)$. Due to brevity and since complete proofs of these results will appear in a forthcoming paper, only a few selected proofs will be included here.

In order to be able to use the known coercivity properties of functional $J_0$ (see [1]) an inequality of the type $W_{\theta,A,p}(u) \geq C J_0(u)$ (for some constant $C$) is highly desired. The following theorem provides sufficient conditions on $\theta$ and $A$ assuring such an inequality.

**Theorem 2.1** Let $\theta \in \hat{\mathcal{M}}(\Omega)$ be such that $\frac{1}{\theta} \in L^\infty(\Omega)$, $A : \Omega \to \mathbb{R}^{2\times 2}$ a measurable matrix field such that $\|A^{-1}(x)\|_{\mathcal{F}} \leq 1 \forall x \in \Omega$ and let $J_0, W_{\theta,A,p}$ be the functionals defined in (5) and (4), respectively. Then $J_0(u) \leq \|\frac{1}{\theta}\|_{L^\infty(\Omega)} W_{\theta,A,p}(u)$ for all $u \in \mathcal{M}(\Omega)$.

The following lemma will be of fundamental importance for proving several of the upcoming results.

**Lemma 2.2** The functional $W_{\theta,A,p}$ defined by (4) is weakly lower semicontinuous with respect to the $L^q$ topology, $\forall q \in [1, \infty)$.

**Theorem 2.3** Let $\Omega \subset \mathbb{R}^2$ be a bounded open convex set with Lipschitz boundary, $\mathcal{X} = L^2(\Omega)$, $\mathcal{Y}$ a normed vector space, $T \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$, $v \in \mathcal{Y}$, $\alpha_1, \alpha_2$ positive constants, $p \in [1, \infty)$, $\theta \in \hat{\mathcal{M}}(\Omega)$ such that $\frac{1}{\theta} \in L^1(\Omega)$ and $\frac{1}{\theta} \in L^\infty(\Omega)$, $A : \Omega \to \mathbb{R}^{2\times 2}$ a measurable matrix field such that $\|A^{-1}(x)\|_{\mathcal{F}} \leq 1 \forall x \in \Omega$. Then the functional

$$F_{\theta,A,p}(u) = \|Tu - v\|_2^2 + \alpha_1 \|\sqrt{1-\theta} u\|_{L^2(\Omega)}^2 + \alpha_2 W_{\theta,A,p}(u), \quad u \in L^2(\Omega).$$

(7)

has a unique global minimizer $\hat{u} \in BV(\Omega)$.

**Proof** By virtue of Proposition 4.1 in [13] it is sufficient to show that the functional

$$W(u) = \alpha_1 \|\sqrt{1-\theta} u\|_{L^2(\Omega)}^2 + \alpha_2 W_{\theta,A,p}(u), \quad u \in L^2(\Omega)$$

(8)

is weakly lower semicontinuous and that $W$-bounded sets are relatively weakly compact in $\mathcal{X}$. To prove the weakly lower semicontinuity of $W$ observe that since $\sqrt{1-\theta} \in L^\infty(\Omega)$ one has $\sqrt{1-\theta} u \to \sqrt{1-\theta} u$. The condition $\frac{1}{\theta} \in L^1(\Omega)$ implies the weak lower semicontinuity of $\|\sqrt{1-\theta} \cdot\|_{L^2(\Omega)}^2$. From this and the weak lower semicontinuity of $W_{\theta,A,p}$ in $L^2(\Omega)$ (see Lemma 2.2) we conclude that the functional $W$ is in fact weakly lower semicontinuous.

To prove the relative weak compactness of $W$-bounded sets, let $\{u_n\} \subset L^2(\Omega)$ be such that $W(u_n) \leq c_1 < \infty$, $\forall n$. We want to show that there exist a subsequence $\{u_{n_j}\} \subset \{u_n\}$ and $u \in L^2(\Omega)$ such that $u_{n_j} \rightharpoonup u$. Since $\{W(u_n)\}$ is uniformly bounded, there exists $K < \infty$ such that $\|\sqrt{1-\theta} u_n\|_{L^2(\Omega)} \leq K \forall n$. From this, the hypothesis $\frac{1}{\theta} \in L^1(\Omega)$ and Theorem 2.1 it follows that $\{u_n\}$ is $BV$-bounded. The existence of a global minimizer of functional (7) belonging to $BV(\Omega)$ then follows from the compact embedding of $BV(\Omega) \subset L^2(\Omega)$ ([2], [3]). Finally note that the condition $\frac{1}{\theta} \in L^1(\Omega)$ implies the strict convexity of $F_{\theta,A,p}$ and therefore the uniqueness of the global minimizer. ■
Remark 2.4 Note that if $\theta(x) \equiv 0$, then $W(u) = \|u\|_{L^2(\Omega)}^2$ and $F_{\theta,A,p}$ as defined in (7) is the classical Tikhonov-Phillips functional of order zero while for $\theta(x) \equiv 1$ a pure anisotropic $p-$BV penalty is obtained. Although the hypotheses of Theorem 2.3 clearly excludes the later case, the next theorem provides conditions for existence and uniqueness of a global minimizer of (7) when $\theta(x) \equiv 1$.

Theorem 2.5 (Anisotropic $p-$BV) Let $\Omega \subset \mathbb{R}^2$ be a bounded open convex set with Lipschitz boundary, $X = L^2(\Omega)$, $Y$ a Hilbert space, $T \in \mathcal{L}(X,Y)$, $v \in Y$, $\alpha$ a positive constant, $p \in [1, \infty)$, $A : \Omega \to \mathbb{R}^{2 \times 2}$ a measurable matrix field such that $\|A^{-1}(x)\|_{p^\ast} \leq 1 \forall x \in \Omega$ and $\partial \chi_\Omega \neq 0$. Then the functional

$$F_{A,p}(u) = \|Tu - v\|_Y^2 + \alpha \sup_{\tilde{v} \in \mathcal{V}_{A,p}} \int_\Omega -u \text{div}(A\tilde{v}) \, dx,$$

where $\mathcal{V}_{A,p} = \{ \tilde{v} : \Omega \to \mathbb{R}^2 \text{ such that } (A\tilde{v}) \in C^0_0(\Omega) \text{ and } |\tilde{v}(x)|_{p^\ast} \leq 1 \forall x \in \Omega \}$, has a global minimizer $\hat{u} \in L^2(\Omega)$. Moreover, if $T$ is injective then the global minimizer is unique.

It is timely to note that in Theorem 2.3 the function $\theta$ cannot assume the extreme values 0 and 1 on a set of positive measure. In some cases a pure anisotropic BV regularization in some regions and a pure $L^2$ regularization in others may be desired, and therefore such a constraint on the function $\theta$ will turn out to be inappropriate. In the next three theorems we introduce different conditions which allow the function $\theta$ to take the extreme values on sets of positive measure.

Theorem 2.6 Let $\Omega \subset \mathbb{R}^2$ be a bounded open convex set with Lipschitz boundary, $X = L^2(\Omega)$, $Y$ a Hilbert space, $T \in \mathcal{L}(X,Y)$, $v \in Y$, $\alpha_1$, $\alpha_2$ positive constants, $p \in [1, \infty)$, $\theta \in \overline{\mathcal{M}}(\Omega)$ and $\Omega_0 = \{ x \in \Omega \text{ such that } \theta(x) = 0 \}$. If $\frac{1}{p} \in L^\infty(\Omega_0^c)$ and $\frac{1}{1-p} \in L^1(\Omega_0^c)$ and $A : \Omega \to \mathbb{R}^{2 \times 2}$ a measurable matrix field such that $\|A^{-1}(x)\|_{p^\ast} \leq 1 \forall x \in \Omega_0^c$. Then functional (7) has a unique global minimizer $u^* \in L^2(\Omega) \cap BV(\Omega_0^c)$.

Theorem 2.7 Let $\Omega \subset \mathbb{R}^2$ be a bounded open convex set with Lipschitz boundary, $X = L^2(\Omega)$, $Y$ a Hilbert space, $T \in \mathcal{L}(X,Y)$, $v \in Y$, $\alpha_1$, $\alpha_2$ positive constants, $p \in [1, \infty)$. Let $\theta \in \overline{\mathcal{M}}(\Omega)$ and $\Omega_1 = \{ x \in \Omega \text{ such that } \theta(x) = 1 \}$. If $\frac{1}{p} \in L^\infty(\Omega_1^c)$, $\frac{1}{1-p} \in L^1(\Omega_1^c)$, $A : \Omega \to \mathbb{R}^{2 \times 2}$ a measurable matrix field such that $\|A^{-1}(x)\|_{p^\ast} \leq 1 \forall x \in \Omega_1^c$ and $\partial \chi_{\Omega_1} \neq 0$, then the functional (7) has a global minimizer $u^* \in L^2(\Omega) \cap BV(\Omega_1^c)$. If moreover $u \in \mathcal{N}(T)$ and $u \neq 0$ implies $u|_{\Omega_1} \neq 0$, then such a global minimizer is unique.

Theorem 2.8 Let $\Omega$, $\chi$, $Y$, $T$, $v$, $\alpha_1$, $\alpha_2$, $p$, $A$ and $\Omega_1$ as in Theorem 2.7, $\Omega_0$ as in Theorem 2.6 and $\theta \in \overline{\mathcal{M}}(\Omega)$. If $\frac{1}{p} \in L^\infty(\Omega_0^c)$, $\frac{1}{1-p} \in L^\infty(\Omega_1^c)$ and $\partial \chi_{\Omega_1} \neq 0$ then the functional (7) has a global minimizer $u^* \in L^2(\Omega) \cap BV(\Omega_1^c \cap \Omega_0^c)$. If moreover $u \in \mathcal{N}(T)$ and $u \neq 0$ implies $u|_{\Omega_1} \neq 0$, then such a global minimizer is unique.
Finally, we present a weak stability result for the minimizers of the functional (7) under different types of perturbations (data, model and regularization parameter).

**Theorem 2.9** Let \( \Omega \subset \mathbb{R}^2 \) be a bounded open convex set with Lipschitz boundary, \( \mathcal{X} = L^2(\Omega) \), \( \mathcal{Y} \) a Hilbert space, \( T, T_n \in \mathcal{L}(\mathcal{X}, \mathcal{Y}), v, v_n \in \mathcal{Y}, \alpha^{(1)}, \alpha^{(2)}, \alpha^{(1)}_n, \alpha^{(2)}_n \) positive constants, for \( n = 1, 2, \ldots \), such that as \( n \to \infty \), \( \alpha^{(1)}_n \to \alpha^{(1)} \), \( \alpha^{(2)}_n \to \alpha^{(2)} \), \( v_n \to v \) and \( T_n u \to Tu \) uniformly for \( u \) in \( W \)-bounded sets where \( W \) is an (8). Let \( p \in [1, \infty), \theta \in \hat{M}(\Omega) \) such that \( 1 - \theta \in L^1(\Omega) \) and \( \frac{1}{\theta} \in L^\infty(\Omega) \), \( A : \Omega \to \mathbb{R}^{2 \times 2} \) a measurable matrix field such that \( \|A^{-1}(x)\|_{p^*} \leq 1 \forall x \in \Omega \). Let

\[
F_{\theta, A, p}(u) = \|Tu - v\|^2_\mathcal{Y} + \alpha^{(1)} \|\sqrt{1 - \theta} u\|^2_{L^2(\Omega)} + \alpha^{(2)} W_{\theta, A, p}(u), \quad u \in L^2(\Omega)
\]

and

\[
F_{(\theta, A, p)_n}(u) = \|T_n u - v_n\|^2_\mathcal{Y} + \alpha^{(1)}_n \|\sqrt{1 - \theta} u\|^2_{L^2(\Omega)} + \alpha^{(2)}_n W_{\theta, A, p}(u), \quad u \in L^2(\Omega).
\]

If \( u, u_n \) are the global minimizers of \( F_{\theta, A, p} \) and \( F_{(\theta, A, p)_n} \), respectively, then \( u_n \xrightarrow{w} u \).

**Proof** The proof of this theorem follows easily from Corollary 3.4 in [8].

3 IMAGE RESTORATION WITH BV-L2 REGULARIZATION

The purpose of this section is to present an application to an image restoration problem of the mixed BV-L2 methods. We will show how those combined methods outperform the pure single ones. For this example, the forward blurring model is given by convolution with a point spread function of “atmospheric turbulence” type, i.e., a two-dimensional Gaussian function with horizontal and vertical standard deviations both equal to 0.02. Data for the inverse problem is then obtained by adding to the blurred image, a zero-mean Gaussian noise with 2% standard deviation.

Figure 1 (a) shows the blurred noisy image while Figure 1 (b) depicts the restoration obtained with a classical order zero Tikhonov-Phillips method (pure \( L^2 \) penalizer). This restoration was later used to build the anisotropic penalization matrix field \( A \) and the weighting function \( \theta \) needed for the mixed BV-L2 regularization. Figure 2 shows the restorations obtained using pure BV penalizers: isotropic in (a) and anisotropic in (b). The better performance of the anisotropic BV method with respect to the isotropic one is clearly seen, particularly in regard to edge detection. This is also reflected in the ISNR values shown in Table 1. Figure 3 shows the restorations obtained with the new mixed BV-L2 penalizer: isotropic in (a) and anisotropic in (b). In both cases, the weighting function \( \theta \) was constructed by computing the normalized convolution of a zero-mean Gaussian function with standard deviation 0.02 and the modulus of the gradient of the regularized solution obtained with a pure zero-order Tikhonov-Phillips method. The construction of the anisotropic penalization matrix field \( A \) was made following the steps in [4].

1150
Gisela L. Mazzieri, Ruben D. Spies and Karina G. Temperini

Figure 1: (a) Blurred noisy image (observation); (b) Tikhonov-Phillips restoration.

Figure 2: (a) Isotropic $BV$ restoration; (b) Anisotropic $BV$ restoration.

The ISNR value (Table 1) shows an improvement of the anisotropic combined method with respect to any to the other methods. It is timely to note that the inclusion of regularity information about the exact solution by means of the anisotropy penalization matrix field $A$, yields better results than the combined isotropic method. The original image is presented in Figure 4.
Figure 3: (a) Mixed $L^2$-isotropic $BV$ restoration; (b) Mixed $L^2$-anisotropic $BV$ restoration.

Figure 4: Original image

Table 1: ISNR values for each restoration

<table>
<thead>
<tr>
<th>Regularization Method</th>
<th>ISNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tikhonov-Phillips</td>
<td>3.806</td>
</tr>
<tr>
<td>Isotropic BV</td>
<td>4.699</td>
</tr>
<tr>
<td>Anisotropic BV</td>
<td>5.132</td>
</tr>
<tr>
<td>Mixed Isotropic</td>
<td>4.639</td>
</tr>
<tr>
<td>Mixed Anisotropic</td>
<td>5.494</td>
</tr>
</tbody>
</table>

4 CONCLUSIONS

In this work we presented several mathematical results on the existence, uniqueness and stability (with respect to perturbations in the data, in the regularization parameter and in
the model) of global minimizers of generalized Tikhonov functionals with penalizers given by convex spatially-adaptive combinations of $L^2$ and anisotropic BV type. An application to image restoration problem was presented. The example showed an improvement of the performance of the new anisotropic combined method with respect to the other methods. We are currently working on further improving several aspects of these combined methods.

ACKNOWLEDGMENTS

This work was supported in part by Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET, through PIP 2014-2016 Nro. 11220130100216-CO, by Universidad Nacional del Litoral, through projects CAI+D PJov 2011 Nro. 50020110100055, CAI+D PI 2011 Nro. 50120110100294 and by the Air Force Office of Scientific Research, AFOSR/SOARD, through Grant FA9550-14-1-0130.

REFERENCES


A METHOD FOR ATOMS SELECTION APPLIED TO SCREENING FOR SLEEP DISORDERS

Román E. Rolon¹, Leandro E. Di Persia¹, Hugo L. Rufiner¹,² and Rubén D. Spies³

¹Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional (sinc(i),UNL-CONICET), Fac. de Ingeniería y Cs. Hídricas, Univ. Nacional del Litoral, Santa Fe, Argentina.
Ciudad Universitaria CC 217, Ruta Nac. 168, km. 472.4 (3000), Santa Fe Argentina
E-mails: rrolon@santafe-conicet.gov.ar, ldipersia@fich.unl.edu.ar and lrufiner@gmail.com

²Laboratorio de Cibernética, Fac. de Ingeniería, Univ. Nacional de Entre Ríos, Argentina.

³Instituto de Matemática Aplicada del Litoral (IMAL), Santa Fe, Argentina.
Colectora Ruta Nac. 168, Km. 0 (3000), Paraje El Pozo, Santa Fe, Argentina
E-mail: rspies@santafe-conicet.gov.ar

Key words: Sparse Representations, Apnea-hypopnea, Inverse problems

Abstract. The Obstructive Sleep Apnea-Hypopnea Syndrome (OSAHS) is a sleep disorder which consists in repetitive events of partial or total airflow decrease during sleep. This pathology has a 4% prevalence in the population around the world and, without appropriate treatment, it increases with age. Actually the gold standard for detecting OSAHS is a polysomnography in a sleep laboratory, which consists in the simultaneous measurement of different physiological signals. In the last years several research studies have shown that the pulse oximetry is a very attractive option of screening for OSAHS, since changes in the dynamics of oxygen in the blood stream (SaO₂) can be related with respiratory problems.

In the last fifteen years, many different signal processing techniques were used for building appropriate representations of a certain types of signals. One of these techniques is known as “sparse representation”. The idea behind the method is to represent the involved signal using only a few coefficients in a certain dictionary, previously constructed. In this work the SaO₂ signal is used in order to predict the occurrence of Apnea-Hypopnea (AH) events. First a dictionary is learned by using a statistical method (NOCICA), then a greedy pursuit algorithm is used in order to obtain the activation coefficients. A subset of the most discriminative coefficients is then selected and used as input of a pattern recognition neural network in order to classify AH events. The problem for finding the optimal dictionary and activation coefficients gives rise to an inverse problem with sparse
constraints. A multilayer perceptron with different number of inputs and neurons in its hidden layer is then tested and the optimal configuration is derived.

1 INTRODUCTION

The American Academy of Sleep Medicine (AASM) distinguishes more than 80 different sleep disorders [1]. One of those pathologies is the Obstructive Sleep Apnea-Hypopnea Syndrome (HS). This pathology is characterized by repetitive episodes of airway narrowing or collapse during sleep. Actually, this pathology affects between 2% to 4% of the population around the world and this prevalence increases with age [2] [3] [4].

The current diagnostic tool for detecting OSAHS is an overnight polysomnography (PSG) in a sleep laboratory whose accessibility in Argentina like in others countries is very limited. Pulse oximetry has become a very attractive option for detecting OSAHS, because its easy accessibility, cost effective and noninvasive and since a decrease in blood oxygen saturation (SaO$_2$) indicates respiratory problems. There is sufficient evidence that if this pathology is not treated properly, the OSAHS is directly related to arterial hypertension, cardiovascular and cerebral vascular diseases, and also high risk of traffic accidents [5], [6].

To grade the OSAHS severity, an index called Apnea Hypopnea Index (AHI) is defined. AHI index is obtained by counting the total number of Apnea-Hypopnea (AH) events per hour while sleeping. A patient with an AHI less than 5 is considered normal, between 5 and 15 mild, between 15 and 30 moderate and more than 30 severe [3].

In the last fifteen years, many different approaches to traditional signal processing problems were taken. Some of these new formulations gave rise to techniques based on non-linear systems and higher-order statistics, including Independent Component Analysis (ICA) [7] and methods to obtain a Sparse Representation (SR) [8] [9] of a signal. They provide new ways of phrasing the problems of signal modeling and representation. One underlying idea is that of representing the involved signals using only a few significant characteristics, e.g. as an SR with just a few basic waveforms.

In a previous work a Most Discriminative Atom Selection (MDAS) [10] method was used in order to detect AH events by using only the SaO$_2$ signal. Now MDAS method is used for screening patients clinically suspected to having a moderate or severe OSAHS.

In this work we start by comparing the performances of an overcomplete assembled dictionary (OAD), trained using class information and a complete dictionary (CD), trained without class information. Those dictionaries were used as generators of an SR of the SaO$_2$ signal, preserving as much as possible the morphology of the signal. After that, the most discriminative activation coefficients, in certain sense, are selected and used to train a Multilayer Perceptron Neural Network (MLP) in order to detect the respiratory events.

The next section describes methods used to learn a dictionary and to inference the coefficients of an SR for a given signal. Preliminary results for diagnosing moderate or
severe OSAHS are presented.

2 METHODS

2.1 SPARSE REPRESENTATION OF SIGNALS

By a dictionary we shall mean a matrix \( \Phi \in \mathbb{R}^{N \times M} \) (with \( M \geq N \)) whose columns \( \phi_j \) are called atoms. A representation of a signal \( s \in \mathbb{R}^N \) in terms of a fixed dictionary \( \Phi \) can be stated as follow:

\[
s = \sum_{i=1}^{M} \phi_j a_j = \Phi a,
\]

where \( a = (a_j) \in \mathbb{R}^M \). Although inappropriately, the term “basis” instead of “dictionary” is sometimes used. Since the atoms are not required to be linearly independent and quite often more atoms than the space dimension are used, the latter situation is usually preferred.

When there are more columns in the dictionary than the size of \( s \), i.e. when \( M > N \) (referred to as an overcomplete dictionary), or when the columns do not form a basis, then there may be non-unique representations of a given signal. In this situation a suitable criterion is required to select only one of those representations. In this context, sparsity often refers to the criterion of choosing a representation with few non-zero coefficients. In particular the SR problem of \( s \) can be stated as follows:

\[
a_{SR} = \arg\min_{a \in \mathbb{R}^M} ||a||_0 \text{ subject to } \Phi a = s,
\]

where \( || \cdot ||_0 \) denotes the zero-norm.

It is important to note that although the mapping \( s \rightarrow a \) in the representation (1) is obviously linear, if the dictionary \( \Phi \) is overcomplete, the mapping \( s \rightarrow a_{SR} \) is not necessary linear. Briefly said, under the sparsity condition (2) the mapping signal-to-coefficients may not be linear.

Let us consider now the problem of finding an SR of a given family of signals with respect to a fixed dictionary \( \Phi \), where learning such a dictionary is part of the problem. Clearly one could learn the dictionary using all the signals in the given family. Although this choice of \( \Phi \) will result in optimal sparsity, most likely it will be highly undesired, mainly because of its size and redundancy. It becomes then necessary to find a dictionary that be optimal, in a certain sense, for a given family of signals.

2.2 FEATURE EXTRACTION

In a previous work [10] a method called Most Discriminative Atoms Selection (MDAS) is developed in order to improve the performance of a MLP. First MDAS selects the columns of a given dictionary that are mostly involved (activated) in the \( SaO_2 \) signal recovery. In this way a \( j^{th} \) column of a dictionary has an activation frequency \( \nu_{c_j} \) given
the class $i$, where $\nu^j_{ci}$ make reference to the number of times that the $j^{th}$ column is used for class $i$ signal recovery.

The candidates to be considered as input of the MLP are then those columns of a given dictionary with higher absolute difference between frequency activation for each of the classes. That is, if some column is active many times for signals with AH events than for the signals without AH events, it is taken into account. Eq.3 shows the absolute difference of activation frequency $F$ given two classes.

$$F = |\nu^j_{c1} - \nu^j_{c2}|$$

Finally MDAS method evaluate the performance of a MLP by varying its number of inputs and neurons in its hidden layer, next an optimal number of inputs $P$ and neurons in hidden layer $Q$ are selected.

### 2.3 Inference about the Activations Coefficients

Now a more general model is taken into account, where the observed signal vectors are noisy linear mixtures of $M$ atoms of a dictionary. Thus, we have:

$$s = \sum_{i=1}^{M} \phi_i a_i + \varepsilon_i = \Phi a + \varepsilon.$$  \hspace{1cm} (4)

Eq. 4 is associated to as a generative model, where a signal vector $s$ is generated from a set of hidden sources $a_j$ aligned as a representation vector $a$, using a fixed dictionary $\Phi$. The elements of $\varepsilon$ are assumed to be uncorrelated, the analysis is then realized in terms of the noise covariance matrix $E\{\varepsilon^T \varepsilon\} = \Lambda_{\varepsilon}^{-1}$, with $E\{\varepsilon\}$ denoting the expected value. The noise vector $\varepsilon$ is then defined by $\varepsilon = s - \Phi a$ and the probability of $s$ given the dictionary $\Phi$ and the hidden sources $a$ can be stated as follows:

$$\pi(s|\Phi, a) = \exp(-\frac{\lambda}{2}|s - \Phi a|^2),$$  \hspace{1cm} (5)

where $\lambda = 1/\sigma^2$ and $\sigma$ is the standard deviation of the additive noise.

The elements of $a$ are assumed to be statistically independent with a joint priori density:

$$\pi_{prior}(a) = \prod_{j=1}^{M} \pi(a_j)$$  \hspace{1cm} (6)

If $\Phi$ is known and $s$ is given, the vector $a$ can be estimated via the Bayes’s formula by considering the posterior distribution (Eq. 7) and a maximum-a-posteriori (MAP) estimation of $a$ is given by Eq. 8

$$\pi_{post}(a|\Phi, s) = \frac{\pi(s|\Phi, a)\pi_{prior}(a)}{\pi(s|\Phi)} \propto \pi(s|\Phi, a)\pi_{prior}(a)$$  \hspace{1cm} (7)
\[ a_{MAP} = \arg\max_a [\pi_{post}(a|\Phi, s)] = \arg\max_a [\log \pi(s|\Phi, a) + \log \pi_{prior}(a)] \] (8)

Lewicki and Olshausen [11] proposed an a priori distribution of Laplacian type, \( \pi(a_j) = N(0, \rho_j) \), where \( \rho_j < 0 \) is given and this leads to the following rule for updating \( a \):

\[ \Delta a = \Phi^T \Lambda \varepsilon - \rho^T |a|, \] (9)

where \( \rho = (\rho_1, \rho_2, ..., \rho_n)^T \).

2.4 DICTIONARY LEARNING

In the previous subsection, a statistical method was used to solve the inference problem about activation coefficients. In what follows the learning problem is similarly solved. The atoms of \( \Phi \) can be estimated by maximizing the log-likelihood function of the data given \( \Phi \) [11], \( L(s, \Phi) = \mathbb{E}[\log \pi(s|\Phi)] \), i.e. as follows:

\[ \hat{\Phi} = \arg\max_\Phi [L(s, \Phi)]. \] (10)

The log-likelihood function can be found by marginalizing the product of the conditional distribution of the data given the dictionary and the prior distribution of the coefficients. That is: \( \pi(s|\Phi) = \int_{R^M} \pi(s|\Phi, a) \pi_{prior}(a) \, da \).

The maximum in Eq. (10) can also be approximated by using a gradient ascent method with the following updating rule [12]:

\[ \Delta \Phi = \eta \Lambda \varepsilon E[(s - \Phi a) a^T |a|]_{\pi_{post}(a|\Phi, s)}, \] (11)

where \( \eta \in (0, 1) \) is the learning coefficient.

Although Eq. (11) provides an explicit updating rule for approximating the solution of (10), the RHS of (11) involves calculating the improper definite integral \( \int_{R^M} (s - \Phi a)^T \pi_{post}(a|\Phi, s) \, da \), which is almost always analytically impossible, and hence, it must be numerically approximated. As the dimension of \( a \) increases, however, this becomes a serious computational challenge and all traditional numerical methods turn out to be inadequate.

Lewicki and Sejnowski [13] used a multivariate Gaussian approximation to the posterior distribution around its maximum \( a_{MAP} \), \( \pi_{post}(a|\Phi, s) \approx \sqrt{\frac{|H|}{2\pi^M}} e^{-1/2(a-a_{MAP})^T H(a-a_{MAP})} \), where \( H \) is the log-posterior evaluated at \( a_{MAP} \). Clearly the mean and covariance matrix of this Gaussian approximation are \( a_{MAP} \) and \( H^{-1} \), respectively. With this approximation the updating rule (11) becomes:

\[ \Delta \Phi = \eta \Lambda \varepsilon ((s - \Phi a)a^T - \Phi H^{-1}). \] (12)

In order to obtain \( \Phi \) and the coefficients (Eqs. (9) and (12)), the implementation proposed by Lewicki and Olshausen [11] was used at the dictionary training stage.
2.5 INVERSE PROBLEMS

The problem for finding the activation coefficients \( \mathbf{a} \), given \( \Phi \) and \( \mathbf{s} \), gives rise to an Inverse Problem (IP) with sparse constraints. A greedy algorithm called Orthogonal Matching Pursuit (OMP) is chosen in order to find a solution of IP. The reason for choosing this algorithm is because it provides a good sparse approximated solution faster than most other methods to solve (2).

The OMP algorithm proposes a modification of the greedy Matching Pursuit algorithm (MP) of Mallat and Zhang \[14\], where the full backward orthogonality of the residual convergence is maintained. It is shown that all additional computation required for the OMP algorithm may be performed recursively \[15\].

Mallat and Zhang show a sequence of approximations used for the MP algorithm in order to obtain an SR of the signal \( \mathbf{s} \). It is assumed that \( \mathbf{a} \) has only \( m \) non-zero components, and therefore the signal vector \( \mathbf{s} = \Phi \mathbf{a} \) is a linear combination of \( m \) columns from \( \Phi \). To identify one such a vector \( \mathbf{a} \), it is necessary to distinguish which columns of \( \Phi \) participate in the measured signal \( \mathbf{s} \). The idea of the OMP algorithm basically consists in properly selecting columns of \( \Phi \). At each iteration, the column most correlated with the current residual of \( \mathbf{s} \) is taken. Then the residual is updated and iterated. In this way, after \( m \) iterations the algorithm will choose a set of \( m \) columns of \( \Phi \) \[16\].

3 EXPERIMENTS

The Sleep Heart Health Study (SHHS) database \(^1\) is used for this work. This database contains exhaustive information about detailed studies which are appropriately designed to investigate the relationships between sleep breathing disorders and cardiovascular diseases. The full dataset contains nearly 1000 complete PSGs, each one of them containing several biomedical signals such as ECG, nasal airflow, respiratory effort and \( SaO_2 \), among others. Annotations of sleep stages, arousals and respiratory events (apnea and hypopnea) are also included. Only the \( SaO_2 \) signals and the AH events will be of our interest.

First of all, a wavelet processing technique is used for denoising the \( SaO_2 \) signal, which was sampled at 1Hz. The denoised \( SaO_2 \) signal is then obtained by making zero the approximation coefficients, at level 8, of the discrete Dyadic Wavelet Transform (DWT) with a mother function Daubechies 2 \[17\]. In the sequel, the \( SaO_2 \) signal shall always refer to the denoised one.

For this work, a subset of 84 studies are selected in order to analyze the performance of a screening method for sleep disorders \[10\]. The dataset is divided into training and test sets consisting of 20 and 64 studies, respectively. The training set contains 4 groups of 5 PSGs corresponding to AHI values below 5, between 5 and 10, between 10 and 15 and above 15. The test set comprises 64 PSGs with different degrees of illness. For both training and test sets, each \( SaO_2 \) signal is segmented into vectors of length 128, with an overlapping of 32 elements. The segments are then arranged as column vectors \( \mathbf{s}_j \in \mathbb{R}^{128} \).

\(^1\)Database: http://physionet.org/physiobank/database/shhpsgdb/
These segments are also labeled as belonging to class 1 or class 2 depending on whether they contain AH events or not, respectively.

Next, we construct two class-training matrices $S_{c1}^{\text{train}}$ and $S_{c2}^{\text{train}}$ by stacking side-by-side all vectors $s$ labeled as class 1 and class 2 in the training set, respectively. The matrix $S_{\text{train}}$ is then defined as $S_{\text{train}} = [S_{c1}^{\text{train}} S_{c2}^{\text{train}}]$, while the matrix $S_{\text{test}}$ is built by stacking side-by-side all vectors $s$ in the test set.

In the next step two complete dictionaries $\Phi_{c1}$ and $\Phi_{c2}$ are learned by using matrices $S_{c1}^{\text{train}}$ and $S_{c2}^{\text{train}}$, respectively. An overcomplete dictionary $\Phi_1$ is then built as $\Phi_1 = [\Phi_{c1} \Phi_{c2}]$. Also, a complete dictionary $\Phi_2$ is learned without taking into account class information, i.e. by using the whole matrix $S_{\text{train}}$. In all cases the learning process is performed by means of the Noise Overcomplete ICA (NOCICA) [13] method.

Now given $s \in \mathbb{R}^M$, a constant $k \in \mathbb{N}$ and a dictionary $\Phi$, we formulate the following linear inverse problem with sparse constraint:

$$ J_{\Phi,s}(v) = \|s - \Phi v\|^2_2 $$

$$ v(\Phi, s, k) = \arg\min_{v \in \mathbb{R}^M, ||a||_0 \leq k} J_{\Phi,s}(v). \quad (13) $$

We denote by $c_1 = v(\Phi_1, s_j, k)$ and $c_2 = v(\Phi_2, s_j, k)$ the solutions of (13) for $s = s_j$, where $s_j$ is the $j^{th}$ column of $S_{\text{train}}$. In the same way, we denote by $d_1 = v(\Phi_1, s_j, k)$ and $d_2 = v(\Phi_2, s_j, k)$ the solutions of (13) for $s = s_j$, where $s_j$ is the $j^{th}$ column of $S_{\text{test}}$.

Finally, for the classification step, the MDAS method described in Section 2.2 is then used in order to select the optimal number of the MLP inputs and neurons in its hidden layer for both dictionaries $\Phi_1$ and $\Phi_2$.

Two classification performance measures are utilized to compare the use of both complete ($\Phi_2$) and overcomplete ($\Phi_1$) dictionaries. For AH events detection, the sensitivity (SE) is defined as the proportion of segments with AH events for whom an event is present, while specificity (SP) is defined as the proportion of segments without AH events for whom an event is not present. Now for moderate or severe OSAHS diagnoses, the sensitivity is defined as the proportion of patients with disease for whom the test is positive, while specificity is defined as the proportion of patients without disease for whom the test is negative. Also, a Receiver Operating Characteristics (ROC) [18] analysis is made, which the following parameters are obtained. True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), cut-off point and Area Under the Curve (AUC).

The detection of patients with a moderate or severe OSAHS is clinically very important. For this reason, in this work an AHI of 15 or more shall be considered as having a positive diagnostic for OSAHS.

4 RESULTS

In the results presented below we used a sparsity level $k = 16$. Also, for improving the performance of the classifiers, we took the same number of segments in each one of both classes. For all 4 methods the optimization problem (13) was solved via OMP.
algorithm due to its high efficiency and effectiveness. Note that the OMP algorithm has to go through 16 iterations to calculate the solution of (13) with $\|a\|_0 = 16$.

In a previous work [10], the performance of an MLP was analyzed by varying its number of inputs and its number of neurons in hidden layer from 2 to 100 and from 10 to 20, respectively. Now, both full vectors $c_1$ and $c_2$ are taken into account for training the MLP (Table 1). The MLP outputs were labeled as “1” or “0” depending on whether an AH event was detected or not, respectively. Finally the estimated AHI (AH\text{est}) was obtained by the total number of “1” divided by the time duration of each study (in hours).

Table 1: Number of inputs and neurons in hidden layer used for training the MLP.

<table>
<thead>
<tr>
<th>Method</th>
<th># inputs</th>
<th># neurons in hidden layer</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDAS-OAD</td>
<td>24</td>
<td>14</td>
</tr>
<tr>
<td>MDAS-CD</td>
<td>30</td>
<td>14</td>
</tr>
<tr>
<td>OAD</td>
<td>256</td>
<td>14</td>
</tr>
<tr>
<td>CD</td>
<td>128</td>
<td>14</td>
</tr>
</tbody>
</table>

Next the whole vectors $d_1$ and $d_2$ were taking into account in order to analyze the performance of the MLP for both overcomplete and complete cases, respectively. Table 2 shows the sensitivity and specificity measures taking account the AH events classifications as well as the corresponding AH\text{est}-AHI correlation percentages. As seen in Table 2, significantly high correlation percentages for detecting AH events were obtained. Among the 64 studies contained in the test set, 48 patients without presence of moderate or severe OSAHS were diagnosed by the expert while 16 studies for whom moderate or severe OSAHS were diagnosed by the expert.

The ROC curves for the diagnostic accuracy for moderate or severe OSAHS detection when using MDAS method are shown in Fig. 1, while Fig. 2 shows the diagnostic accuracy for moderate or severe OSAHS without applying MDAS method.

Table 2: Classifier’s performance for AH events detection

<table>
<thead>
<tr>
<th>Method</th>
<th>SE</th>
<th>SP</th>
<th>Correlation percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDAS-OAD</td>
<td>74.52%</td>
<td>76.73%</td>
<td>90.04%</td>
</tr>
<tr>
<td>MDAS-CD</td>
<td>68.86%</td>
<td>67.69%</td>
<td>74.57%</td>
</tr>
<tr>
<td>OAD</td>
<td>62.84%</td>
<td>65.63%</td>
<td>91.24%</td>
</tr>
<tr>
<td>CD</td>
<td>61.13%</td>
<td>64.85%</td>
<td>92.23%</td>
</tr>
</tbody>
</table>

Table 3 shows the values of the performance measures for OSAHS detection obtained for all 4 methods. First with the MDAS-OAD method, a total of 15 patients were correctly detected as having OSAHS and 44 patients were also correctly diagnosed without
presence of OSAHS. Moreover this method resulted in 4 false positive and 1 false negative
diagnoses. With the MDAS-CD method, all 16 patients with OSAHS were detected and
43 patients were correctly diagnosed without presence of OSAHS. This method resulted in
5 false positives and no false negative diagnoses. Analogously, the performance measures
obtained with the OAD and CD methods are shown in the last two rows of Table 3.

Looking at the last column the Table 3 we see the AUC values obtained with all
4 methods are significantly high, corresponding the highest value to the CD method
(AUC=0.9939). Table 3 also shows that although the application of the MDAS method
has a negative impact in the specificity measure, it has no effect on the sensibility. Sim-
ilarly the specificity for MDAS-OAD and OAD methods resulted higher than specificity
obtained with MDAS-CD and CD methods, respectively.
Table 3: Performance measures for diagnosing moderate or severe OSAHS.

<table>
<thead>
<tr>
<th>Method</th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>SE</th>
<th>SP</th>
<th>Cut-off</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDAS-OAD</td>
<td>15</td>
<td>44</td>
<td>4</td>
<td>1</td>
<td>93.75%</td>
<td>91.67%</td>
<td>399.65</td>
<td>0.9799</td>
</tr>
<tr>
<td>MDAS-CD</td>
<td>16</td>
<td>43</td>
<td>5</td>
<td>0</td>
<td>100%</td>
<td>89.58%</td>
<td>5.8382</td>
<td>0.8958</td>
</tr>
<tr>
<td>OAD</td>
<td>15</td>
<td>46</td>
<td>2</td>
<td>1</td>
<td>93.75%</td>
<td>97.92%</td>
<td>306.85</td>
<td>0.9899</td>
</tr>
<tr>
<td>CD</td>
<td>16</td>
<td>46</td>
<td>2</td>
<td>0</td>
<td>100%</td>
<td>95.83%</td>
<td>6.66</td>
<td>0.9939</td>
</tr>
</tbody>
</table>

5 CONCLUSIONS

The previous analysis shows that the sparse representation of an $SaO_2$ signal is a suitable technique for detecting moderate or severe OSAHS. The algorithm NOCICA was found to be a very useful tool for learning both overcomplete and complete dictionaries. Also, to linearly regress the apnea-hypopnea events a multilayer perceptron was successfully constructed.

The obtained results show that although the sensibility and specificity values (Table 2) were not so good, a high correlation between the AHI observed by the expert via PSG and the AHIest obtained by using the sparse representation of the $SaO_2$ signal. This fact constitutes a strong evidence that such a procedure could be successfully used for detecting moderate or severe OSAHS. Also, as shows Table 3 a very high performance for moderate or severe OSAHS diagnosing was achieved.

Although similar classification results were obtained with and without applying the MDAS method, it is important to point out that the use of this procedure reduces highly the dimensionality of the problem, and therefore the CPU time required for classification.

As mentioned in Section 4, a sparsity level of $k = 16$ was selected. For future works we propose to vary $k$ from a low to a high level and analyze its impact on the diagnostic performance measures. Also, the number of inputs $P$ and the threshold of the MLP output will be further studied.

6 ACKNOWLEDGEMENTS

This work was supported in part by Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET, through PIP 2014-2016 Nro. 11220130100216-CO and PIP 2012-2014 Nro. 114 20110100284-KA4, by the Air Force Office of Scientific Research, AFOSR / SOARD, through Grant FA9550-14-1-0130 and by Universidad Nacional del Litoral through project CAI+D 501 201101 00519 LI.

REFERENCES


AD-HOC GAUSSIAN DICTIONARIES FOR SPARSE REPRESENTATION OF EVOKED RELATED POTENTIALS

VICTORIA PETERSON\textsuperscript{1}, HUGO L. RUFINER\textsuperscript{2} AND RUBEN D. SPIES\textsuperscript{3}

\textsuperscript{1}Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral, Consejo Nacional de Investigaciones Científicas y Técnicas (sinc(i)-FICH-UNL-CONICET).
Ruta Nac. No 168, km 472.4 (3000), Santa Fe - Argentina
vpeterson@santafe-conicet.gov.ar

\textsuperscript{2}Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral, Consejo Nacional de Investigaciones Científicas y Técnicas (sinc(i)-FICH-UNL-CONICET)
Ruta Nac. No 168, km 472.4 (3000), Santa Fe - Argentina and Facultad de Ingeniería, Universidad Nacional de Entre Ríos (FI-UNER) Ruta Prov. No 11, km 10 (3100), Oro Verde - Entre Ríos - Argentina
lrufiner@fich.unl.edu.ar

\textsuperscript{3}Instituto de Matemática Aplicada del Litoral (IMAL-CONICET-UNL)
Ruta Nac. No 168, Paraje El Pozo (3000), Santa Fe - Argentina and Facultad de Ingeniería Química, Universidad Nacional del Litoral, Consejo Nacional de Investigaciones Científicas y Técnicas (FIQ-UNL)
Santiago del Estero 2829 (3000), Santa Fe - Argentina.
rspies@santafe-conicet.gov.ar

Key words: Brain-Computer Interface, Evoked Related Potentials, Sparse Representation, $\ell_1$-minimization, Robust Classification

Abstract. A Brain-Computer Interface (BCI) is a system which provides direct communication between the mind of a person and the outside world by using only brain activity (EEG). A common EEG BCI paradigm is based on the so called Event-Related Potentials (ERP) which are responses of the brain to some external stimuli. For the present work at hand, the innermost part of a BCI is the pattern recognition stage whose aim is to detect the presence of ERPs with high accuracy. In recent years there has been a growing interest in the study of sparse representation of signals. Using a dictionary composed of prototype atoms, signals are written as linear combinations of just a few of those atoms. This sparse representation is found to be appropriate for posterior classification purposes. In this work we propose a sparse representation and posterior classification of ERPs signals by means of an ad-hoc spatio-temporal dictionary composed of bidimensional Gaussian atoms. Based on $\ell_1$-minimization we find the sparsest possible solution which allow us to design a robust classification based on nearest representation
INTRODUCTION

Sparse representations have received great interest in the recent years due to their success in many applications in signal and image processing [1], [2].

Given a \( m \times n \) matrix \( A \), called dictionary, an unknown signal \( x \in \mathbb{R}^n \) and a measurement vector \( y \in \mathbb{R}^m \), we seek to find the sparsest coefficient vector \( x \) such that
\[
y = Ax.
\]
Recovering \( x \) given \( A \) and \( y \) is a non-trivial inversion problem since in general the size of \( x \) is greater than the size of \( y \), the measurement \( y \) is contaminated by noise and the problem is usually severally ill-posed. Hence, regularization is required.

Given a functional \( J(x) \), which penalizes certain undesired properties of the solution, a regularized solution can then be obtained by solving the following constrained optimization problem:
\[
(P_J) : \min_{x \in \mathbb{R}^n} J(x) \quad s.t. \quad y = Ax.
\]
Most of the existing works on sparse learning are based on variants of the \( \ell_1 \)-norm regularization \( (J(x) = ||x||_1) \) due to its sparsity-inducing property, convenient convexity, wide and strong theoretical support, and great success in several applications [3]. Thus, the regularized sparse solution are obtained by solving the following minimization problem:
\[
(P_1) : \min_{x \in \mathbb{R}^n} ||x||_1 \quad s.t. \quad y = Ax.
\]
It has been shown, under certain conditions, that the minimum \( \ell_1 \)-norm solution of an undetermined linear system is also the sparsest possible solution [4]. In practice, however, \( y \) is contaminated by noise and therefore the equality constraint in (2) must be relaxed:
\[
(P_{1,2}) : \min_{x \in \mathbb{R}^n} ||x||_1 \quad s.t. \quad ||y - Ax||_2 \leq \epsilon,
\]
where \( \epsilon \) is noise level.

Problem \((P_{1,2})\), known as the basis pursuit denoising problem (BPDN), is equivalent to the following unconstrained minimization problem:
\[
(P_\lambda) : \min_{x \in \mathbb{R}^n} \frac{1}{2}||y - Ax||_2^2 + \lambda ||x||_1,
\]
which can be viewed as a generalized Tikhonov-Phillips regularization functional [5] or in a statistical context as a Least Absolute Shrinkage and Selection Operator (LASSO) [6].

In this article we pretend to use the virtues of sparse representation framework together with the ideas proposed in [7] in order to develop a robust classification method in the context of Event Related Potential (ERP) detection.

The problem at hand comes from Brain-Computer Interfaces (BCI) systems. BCI can significantly improve the quality of life of a person who cannot control his/her own body or even is not able to communicate. By using only brain activity, BCI provides a person a new way of communication and control without needing any peripheral nerves or muscles.
The most common and non-invasive method used to decode the intention of a BCI user consists of detecting the presence of ERP signals in electroencephalogram (EEG) records.

In the “oddball” paradigm it is well known that when a person is stimulated with some external and “rare” item (which can be auditory, visual or somatosensory) an ERP is elicited. One of the main components of ERP signals is an enhanced positive-going component with a latency of about 300 ms (called P300 wave) [9], [10]. Unfortunately detecting a P300 wave (which means detecting the ERP signal) is not an easy task. Mainly due to the fact that SNR between ERP and EEG signals is very low (about -50 db) and also due to the large variation present in P300 wave records among different trails.

In order to use all the information that can be found in EEG records we shall construct an ad-hoc dictionary whose elements are composed by two bidimensional Gaussian atoms representing the main spatio-temporal variation of EEG records with and without P300, respectively. As explained in the next section, the dictionary was built based on a certain neurologically well-supported assumptions on the P300 wave.

2 MATERIALS AND METHODS

2.1 Database

An Open-Access P300 speller database was used [11]. EEG records from 18 subjects were acquired by 10 electrodes in the positions shown in Fig.1. The records were digitalized at a rate of 256 Hz.

![Figure 1: Positions of the electrodes used in EEG acquisition. Modified from [11].](image)

The Farwell and Donchin oddball paradigm was used [10], where a 6-by-6 matrix containing letters and numbers were displayed on a computer screen. During the experiment, a subject is asked to spell different words. The person must focus on one character at the time. As stimulus, a row or a column of the matrix is randomly flashed. In each stimulating block, every row and column of the matrix is intensified only once. If the person is well concentrated, when the chosen character is illuminated, a relevant event occurs, i.e. an ERP signal is elicited [12]. Since there are two classes involved, the 6-by-6 matrix result in twelve possible events (one per each row and one per each column), of which only two are relevant, i.e. with ERP, in every intensification block.
Each subject participated in 4 sessions, where the first two ones were copy-spelling runs. For this reason, in this work we used the first and second sessions as training and testing sets, respectively. Each session consisted of 15 trials, resulting in 2880 EEG epochs per channel (480 target records and 2400 non-target records) for the training set and 900 EEG epochs per channel (150 of them being target) for the testing set. An epoch is a EEG record of one second duration extracted at the beginning of an intensification.

2.2 Gaussian Dictionary Generation

2.2.1 Image Pre-processing

In order to use all the information available in an epoch, we used not only the information time-to-time but also the spatial information given by the electrode’s positions. In this way, we constructed one image per trial in the time-channel plane. Since a trial is an epoch of one second duration extracted at the beginning of an intensification, and the number of channels was ten, our images consisted of $256 \times 10$ pixels.

Analyzing the images of the great averages for both classes (target records and non-target records), we noticed that the shape of the P300’s peaks could be significantly improve by re-ordering the channels. In order to do this we proceeded as follows. Since the registered amplitude of the P300 wave differs according to the sensors’s position, the channels were re-ordered based on decreasing signal energy, i.e, we chose as the first channel the one possessing greatest $\ell_2$-norm, and the rest were decreasingly ordered by their Euclidean distance at this “first” channel. Next we filtered the image with a median filter. An example of a resulting re-ordered and filtered image are shown in column 2 of both figures Fig.2a and Fig.2b.

After analyzing each re-ordered and filtered target image, we observed that for all of them there were one or two notable peaks between 0.2s and 0.6s. On the other hand, the peaks for non-target images always came from the oscillatory background. For those reasons and because the 0.2-0.6 range is in agreement with the latency window of the P300 wave, we cropped all images between 0.2s and 0.6s, resulting in images of $104 \times 10$ pixels.

Images belonging to the three different pre-processing stages are shown in Fig.2 for the first subject in the database. There are the original image, the re-ordered and filtered image and the re-ordered, filtered and cropped image for both classes in two different views.

In the sequel, a template will referred to the re-ordered, filtered and cropped images.

2.2.2 Levenberg-Marquardt Estimation of Mother Elements

Inspired by the ideas of pre-defined dictionaries (like wavelet dictionaries), we want to generate a dictionary by means of variations of an appropriate “mother” element.

In our case we shall consider one mother element per class and per subject, given by a
linear affine combination of two Gaussian functions (atoms). More precisely, let
\[ \mathcal{P} = \{ (p_1, p_2, ..., p_{11}) \in \mathbb{R}^{11} : p_5, p_6, p_{10}, p_{11} > 0 \} \] and for \( p \in \mathcal{P} \) define:
\[
z(t, c; p) = p_1 + p_2 \exp \left( \frac{(t - p_3)^2}{p_5^2} + \frac{(c - p_4)^2}{p_6^2} \right) + p_7 \exp \left( \frac{(t - p_8)^2}{p_{10}^2} + \frac{(c - p_9)^2}{p_{11}^2} \right).
\] (5)

In the sequel we shall always identify the scalar field \( z(t, c; p) \) with the vector \( z(p) \in \mathbb{R}^{1040} \) obtained after stacking in a column vector the matrix resulting of the evaluation of \( z(t, c; p) \) over a \( 104 \times 10 \) image grid.

Given the template \( f(t, c) \) (or simply \( f \in \mathbb{R}^{1040} \)) we formulate the corresponding non-linear fitting problem associated to \( f \) and the mother element as:
\[
(F_p) : \min_{p \in \mathcal{P}} \| f - z(p) \|_2^2.
\] (6)

The Levenberg-Marquardt (LM) method [13], [14] is a standard and efficient technique to solve nonlinear least squares problems. We used it to find \( p \) in (6).

The estimated mother elements for the same subject used in Fig.2 are shown in Fig.3. The templates are shown on the left while the two estimated Gaussian elements fitted by LM algorithm are on the right.

It is timely to observe the good fitting of the mother elements to the corresponding templates, especially for the target class.

### 2.2.3 Dictionary Build Up

As the dictionary must capture the main variation of the P300 wave for each subject, in order to construct an appropriate dictionary it is very important to analyze the sensitivity
of each one of the components of the parameters vector \( p \) with respect to the epochs in each one both classes. With this in mind, we performed a rough analysis in this direction following the next steps:

1. **Averaging and balancing**: take averages of 5 trials for target records and averages of 25 trials for non-target records (see Remark (1) below).

2. **Template generation**: for each one of the averaged trials generated in step (1) construct the corresponding template image, as described in Section 2.2.1.

3. **Parameters generation**: for each one of the templates described in step (2) estimate the corresponding parameter \( p \) by solving problem (6) with the LM method.

4. **Parameters range estimation**: step (3) generates a family of vectors \( p \in \mathbb{R}^{11} \) for each one of both classes. The range variation for each one of the 11 components of \( p \) is estimated within each one of both families in the usual way.

**Remark (1).** The different number of elements, 5 and 25, for taking averages in each one of the classes has the objective of balancing the different number of epochs among classes.

For the dictionary generation itself we varied one parameter at the time while the others were kept constant. The variations were made in order to cover the whole range.
in each one of the components of \( p \) in one hundred equal increments\(^1\).

From the sparse representation point of view, it is highly desirable not to have dictionary’s elements which are “too similar”. The Mutual Coherence (MC) of a dictionary, denoted by \( \mu(A) \), measures the similarity between dictionary’s elements. It is defined as the maximal absolute scalar product between two different \( \ell_2 \)-normalized elements of the dictionary \( A \), that is:

\[
\mu(A) = \max_{i \neq j} | \langle a_i, a_j \rangle |. \tag{7}
\]

Notice that if two dictionary’s elements are parallel then \( \mu(A) = 1 \), otherwise \( \mu(A) \geq 0 \).

In the dictionary generation algorithm we discarded an element if its MC was greater than some predefined number \( \kappa \). To avoid classification bias, the dictionary sizes should be the same for both classes. With this objective in mind, we randomly eliminated the necessary number of elements from the larger dictionary whose MCs were greater than another predefined value (e.g. \( \kappa - 0.5 \)).

At this point we have for any subject one dictionary per class, both dictionaries having the same number of elements and with some predefined \( MC = \kappa \). It is appropriate to mention here that the size of the dictionary grows with \( \kappa \).

### 2.3 Classification Based on Sparse Representation

Let us define a new matrix \( A \) as the concatenation of the \( n \) elements from both target and non-target dictionaries, \( A_1 \) and \( A_2 \), respectively, that is:

\[
A = [A_1 \ A_2] = [a_{1,1}, a_{1,2}, \ldots, a_{1,n}, a_{2,1}, \ldots, a_{2,n}]. \tag{8}
\]

For given \( \epsilon, A \) and \( y \), let \( \hat{x} \) be the solution of the problem \((P_{1,2})\) in (3). Ideally one would expect that the nonzero entries of \( \hat{x} \) will correspond to columns of \( A \) belonging all to the same class. In that case the association of \( y \) to one of both classes is clearly trivial. Noise and modeling errors, nonetheless, may lead to nonzero entries associated to the wrong class. It is therefore clear that any classification criterion will have to take into account additional information such as the goodness of fit, that is, a measure of how well the coefficients associated with each one of the classes reproduced \( y \).

For each class \( i \), let \( \delta_i : \mathbb{R}^n \to \mathbb{R}^n \) be the lifting function that selects the coefficients associated with the \( i^{th} \) class, that for \( x \in \mathbb{R}^n, \delta_i(x) \in \mathbb{R}^n \) is the vector whose only nonzero entries are the elements in \( x \) that are associated to class \( i \). The representation of a given test sample \( y \) in class \( i \) is then \( \hat{y}_i = A \delta_i(\hat{x}) \). The classification of \( y \) proceeds by assigning it to the class that minimizes the residual, i.e., we associate to \( y \) the class given by:

\[
\arg \min_{i=1,2} r_i(y) = ||y - A \delta_i(\hat{x})||_2. \tag{9}
\]

The Sparse Representation-based Classification (SRC) algorithm (as proposed in [7]) can then be written as follows:

\(^1\)The variation in \( p_1 \), corresponding to the offset parameter in (5) was negligible and therefore it was kept constant for the dictionary generation.
1. **Input:** The train dictionary $A = [A_1, A_2] \in \mathbb{R}^{m \times n}$, a test sample $y \in \mathbb{R}^m$ and $\epsilon$ tolerance or $\lambda$ regularized parameter.

2. Normalize the columns of $A$ to have unit $\ell_2$-norm.

3. Solve the $\ell_1$-minimization problem:
   \[
   \min_{x \in \mathbb{R}^n} ||x||_1 \quad s.t. \quad ||y - Ax||_2 \leq \epsilon.
   \]
   Or equivalently solve:
   \[
   \min_{x \in \mathbb{R}^n} \frac{1}{2}||y - Ax||_2^2 + \lambda||x||_1.
   \]

4. Compute the residuals $r_i(y) = ||y - A \hat{x}_i||_2$, for $i=1,2$.

5. **Output:** Identity($y$) = $\arg\min_{i=1,2} r_i(y)$.

### 3 RESULTS AND DISCUSSIONS

The sparse vectors $\hat{x}$ for each test observation $y$ were estimated using different functions by the application of SLEP 4.1 toolbox [15]. More precisely we estimated the sparse vectors $\hat{x}$ as follows:

1. $\ell_1$-ball constrained least squares problem (LeastC):
   \[
   \min_{x} \frac{1}{2}||y - Ax||_2^2 \quad s.t. \quad ||x||_1 < \epsilon,
   \]
   where $\epsilon$ is the radius of the $\ell_1$-ball.

2. $\ell_1$-norm regularized least squares problem (LeastR):
   \[
   \min_{x} \frac{1}{2}||y - Ax||_2^2 + \lambda||x||_1.
   \]

3. Non-negative $\ell_1$-ball constrained least squares problem (NNLeastC):
   \[
   \min_{x} \frac{1}{2}||y - Ax||_2^2 \quad s.t. \quad ||x||_1 < \epsilon, x \geq 0.
   \]

4. Non-negative $\ell_1$-norm regularized least squares problem (NNLeastR):
   \[
   \min_{x \geq 0} \frac{1}{2}||y - Ax||_2^2 + \lambda||x||_1.
   \]
The optimal values of $\lambda$ and $\epsilon$ were fixed after analyzing the residue as a function of $\lambda$ or $\epsilon$, respectively. We varied the parameters between $10^{-4}$ and 1 in increments of $10^{-4}$. We finally chose $\lambda = 0.05$ and $\epsilon = 0.8$ as optimal values. The behavior of the residue is shown in Fig.4.

In order to analyze the impact of the choice of MC value on generation of the dictionary and, consequently, on the classification rate obtained by SRC algorithm, we varied the MC values from 0.500 to 0.995 in increments of 0.005.

The unweighted accuracy rate (UAR) was used as the performance classification measure. Fig.5 shows the classification results reached by each optimization problem (LeastC, LeastR, NNLeastC and NNLeastR) for four different database’s subjects.

There does not seem to be a clear tendency between classification results and MC values. Note that while for subject No 10 the UAR shows a clearly increasing tendency for increasing MC values, the opposite happens for subject No 1. This observation constitutes a solid reason for developing a subject depended classification tools.

Table 1 summarizes the best UAR results per each subject with the different optimization problems. The corresponding MC value is shown between parenthesis. An analysis of Table 1 seems to suggest that better classification result are obtained by imposing $x \geq 0$. Although further analysis is required, we strongly believe that this is due to the fact that the elements in the target dictionary present mainly positive peaks.

4 CONCLUSIONS

In this paper, we constructed an ad-hoc Gaussian dictionary for representing the P300 wave in a channel-time space. We focused our work in constructing a suitable dictionary per each subject in order to represent the variation of the P300 wave of that particular subject. It is not of our interest to find a generalized (all subjects) representation of the P300 wave. Moreover, we have good reasons to believe that better classification result can be obtained by improving the representation of non-target signals.
The resulting $\ell_1$-minimization problems were solved using two different optimization approach both with and without a positive constraint, and used the sparse information of the $\hat{x}$ solution for classification purposes. Although the classification performances are far from being optimal, we point out that the classification method used is very simple and moreover noisy single trial epochs were used. Much further research is clearly needed in this regard. In particular it is of great interest to find alternative and/or complementary ways to the $\ell_1$-minimization approach, which could allow classification improvement. We are currently devoting efforts in this direction.

5 ACKNOWLEDGEMENTS

This work was supported in part by Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET, through PIP 2014-2016 Nro. 11220130100216-CO, the Air Force Office of Scientific Research, AFOSR/SOARD, through Grant FA9550-14-1-0130 and by CAID-UNL 2011.

REFERENCES


Table 1: Best UAR results for the different optimization problems.

<table>
<thead>
<tr>
<th>Subject No</th>
<th>LeastC</th>
<th>LeastR</th>
<th>NNLeastC</th>
<th>NNLeastR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.600 (0.565)</td>
<td>0.569 (0.555)</td>
<td>0.614 (0.555)</td>
<td>0.607 (0.555)</td>
</tr>
<tr>
<td>2</td>
<td>0.583 (0.900)</td>
<td>0.579 (0.945)</td>
<td>0.628 (0.650)</td>
<td>0.620 (0.795)</td>
</tr>
<tr>
<td>3</td>
<td>0.531 (0.680)</td>
<td>0.531 (0.730)</td>
<td>0.503 (0.780)</td>
<td>0.509 (0.680)</td>
</tr>
<tr>
<td>4</td>
<td>0.558 (0.970)</td>
<td>0.546 (0.645)</td>
<td>0.583 (0.955)</td>
<td>0.590 (0.965)</td>
</tr>
<tr>
<td>5</td>
<td>0.528 (0.670)</td>
<td>0.540 (0.700)</td>
<td>0.559 (0.570)</td>
<td>0.567 (0.600)</td>
</tr>
<tr>
<td>6</td>
<td>0.547 (0.855)</td>
<td>0.533 (0.750)</td>
<td>0.539 (0.855)</td>
<td>0.538 (0.855)</td>
</tr>
<tr>
<td>7</td>
<td>0.554 (0.645)</td>
<td>0.558 (0.900)</td>
<td>0.545 (0.980)</td>
<td>0.548 (0.995)</td>
</tr>
<tr>
<td>8</td>
<td>0.576 (0.920)</td>
<td>0.561 (0.790)</td>
<td>0.537 (0.995)</td>
<td>0.539 (0.995)</td>
</tr>
<tr>
<td>9</td>
<td>0.531 (0.650)</td>
<td>0.544 (0.595)</td>
<td>0.551 (0.700)</td>
<td>0.604 (0.635)</td>
</tr>
<tr>
<td>10</td>
<td>0.612 (0.795)</td>
<td>0.610 (0.745)</td>
<td>0.663 (0.960)</td>
<td>0.664 (0.985)</td>
</tr>
<tr>
<td>11</td>
<td>0.557 (0.910)</td>
<td>0.565 (0.940)</td>
<td>0.595 (0.995)</td>
<td>0.624 (0.940)</td>
</tr>
<tr>
<td>12</td>
<td>0.519 (0.680)</td>
<td>0.517 (0.600)</td>
<td>0.545 (0.635)</td>
<td>0.547 (0.640)</td>
</tr>
<tr>
<td>13</td>
<td>0.580 (0.865)</td>
<td>0.577 (0.840)</td>
<td>0.639 (0.740)</td>
<td>0.647 (0.625)</td>
</tr>
<tr>
<td>14</td>
<td>0.567 (0.530)</td>
<td>0.586 (0.540)</td>
<td>0.670 (0.535)</td>
<td>0.673 (0.555)</td>
</tr>
<tr>
<td>15</td>
<td>0.573 (0.700)</td>
<td>0.561 (0.645)</td>
<td>0.585 (0.960)</td>
<td>0.559 (0.855)</td>
</tr>
<tr>
<td>16</td>
<td>0.523 (0.585)</td>
<td>0.523 (0.590)</td>
<td>0.614 (0.995)</td>
<td>0.609 (0.995)</td>
</tr>
<tr>
<td>17</td>
<td>0.577 (0.590)</td>
<td>0.581 (0.855)</td>
<td>0.589 (0.995)</td>
<td>0.595 (0.605)</td>
</tr>
<tr>
<td>18</td>
<td>0.546 (0.920)</td>
<td>0.545 (0.760)</td>
<td>0.577 (0.975)</td>
<td>0.577 (0.980)</td>
</tr>
</tbody>
</table>


LOCAL EDGE-PRESERVING IMAGE INPAINTING WITH MIXED WEIGHTED ANISOTROPIC REGULARIZATION METHODS

FRANCISCO J. IBARROLA¹ AND RUBÉN D. SPIES¹,²

¹Departamento de Matemática, Facultad de Ingeniería Química, Universidad Nac. del Litoral
Santiago del Estero 2829, S3000AOM Santa Fe, Argentina
franciscoj.iba@gmail.com

²Instituto de Matemática Aplicada del Litoral, IMAL, CONICET-UNL.
Centro Científico Tecnológico CONICET Santa Fe
Colectora Ruta Nac. 168, km 472, Paraje “El Pozo”, 3000, Santa Fe, Argentina
rspies@santafe-conicet.gov.ar

Key words: Inverse problems, Inpainting, Regularization, Diffusion

Abstract. The image inpainting problem consists of restoring an image from a possibly noisy version of it, in which data from one or more regions is missing. Over the last couple of decades, several inpainting models to perform this task have been developed. Although some of these models perform reasonably well in certain types of images, quite a few issues are yet to be sorted out. For instance, if the true image is smooth, the inpainting can be performed with very good results by means of a Bayesian approach and a maximum a posteriori computation [1]. For non-smooth images, however, such an approach is far from being satisfactory. Even though the introduction of anisotropy by prior gradient inpainting to the latter methodology is shown to produce satisfactory results for slim missing regions, the quality of the restoration decays as the occlusions widen. On the other hand, Total Variation (TV) inpainting models based on high order PDE diffusion equations can be used whenever edge restoration is a priority. More recently, the introduction of spatially variant conductivity coefficients to these models, such as in the case of Curvature-Driven Diffusions (CDD) [4], has allowed inpainted images with well defined edges and enhanced object connectivity. The CDD approach, nonetheless, is not quite suitable wherever the image is smooth, as it tends to produce piecewise constant restorations.

In this work we shall combine the ideas of gradient inpainting for introducing anisotropy along with the notion of isophote curvature into a mixed penalization inpainting model [5] to produce an edge-preserving restoration which also allows for inpainting of smooth areas without precluding object connectivity. Several results will be presented and comparisons will be made in order to illustrate the performance of the different penalizers in conjunction with the corresponding parameter estimation criteria and computing methods.
1 INTRODUCTION

Let us consider an image from which data from one or more regions is missing. The inpainting problem consists of using the available information from the occluded image for filling up the missing areas in a way that looks natural and intuitive to the human eye. Thus, for instance, if an occlusion runs over a smooth area, we would expect the inpainting to be smooth, while if the occlusion goes through an edge, we would expect the two parts set apart by the occlusion to result appropriately connected. These desired properties are often hard to include in a mathematical model, even more so, simultaneously. In this article we shall combine the Curvature Driven Diffusion (CDD) model proposed in [4] with the ideas of including structural information [1] into regularization problems, to develop an inpainting model that can better comply with the aforementioned criteria.

We shall begin by fixing some notation. Our (grayscale) image will be associated to a function \( u : \Omega \subset \mathbb{R}^2 \to [0,1] \), where \( u(x,y) \) represents the light intensity of the point \((x,y)\), corresponding \( u = 1 \) to white, and \( u = 0 \) to black. With \( D \) we shall denote the occluded region and with \( v = u|_{\partial \Omega \setminus D} \) the known part of \( u \).

Let us now take a look at a few known methods appropriate for inpainting certain types of occluded images.

1.1 Tikhonov-Phillips Inpainting Methods

If the image to be inpainted is smooth, then there are traditional methods that turn out to be satisfactory. For instance, if we consider the Tikhonov-Phillips regularization method of order 1 (T1), then the inpainted image is defined as

\[
\hat{u} = \arg \min_{u \in L^2(\Omega)} \{ \| Tu - v \|_{L^2}^2 + \lambda \| \nabla u \|_{L^2}^2 \},
\]

where \( T : L^2(\Omega) \to L^2(\Omega) \) is the \( D \)-occlusion operator, defined by \( Tu = u_{\chi_{\Omega \setminus D}} \), and \( \lambda > 0 \) is a regularization parameter.

Nonetheless, if the original image has edges or borders set apart by an occlusion, then such inpainting methods fail to comply with the so called connectivity principle [4], in the sense that they are unable to appropriately reconnect the broken parts (see Figure 1(b)). There are alternatives to enhance edge preservation in these kind of methods by appropriately modifying the penalizing term in (1). One of such alternatives is the introduction of an anisotropy-inducing matrix field \( A \) [1], which diminishes penalization on the maximum gradient direction. The restored image in this case is given by

\[
\hat{u} = \arg \min_{u \in L^2(\Omega)} \{ \| Tu - v \|_{L^2}^2 + \lambda \| A \nabla u \|_{L^2}^2 \}.
\]

Although this method works significantly better for “thin” occlusions, it becomes inappropriate as the occlusion widens (see Figure 1(c)). We shall then explore other methods possessing better edge preserving properties.
1.2 TV inpainting

One of the most commonly used edge-preserving methods is the so called total variation (TV) regularization [2], [3]. With this method, the inpainted image is given by

$$\hat{u} = \arg \min_{u \in L^2(\Omega)} \{ \| Tu - v \|_{L^2}^2 + \lambda \| \nabla u \|_{L^1} \}. \quad (3)$$

By using the corresponding Euler-Lagrange equation, it can be seen that the minimizer in (3) is the steady state solution of the following diffusion PDE (see [6])

$$\frac{\partial u}{\partial t} = \nabla \cdot \left[ \frac{\nabla u}{|\nabla u|} \right]. \quad (4)$$

This method tends to produce piecewise constant solutions with sharp edges inside the occlusions (this can be seen in Figure 2). Although this inpainting process seems to fairly comply with the aforementioned connectivity principle, it presents a peculiar property. In fact, if we look at the black vertical bar in Figure 2(a), we notice that it is wider than the occlusion’s width. However, if the occlusion is wider than the bar’s width (Figure 3(a)), then the TV solution is very different (see Figure 3(b)). Although there is no formal reason to sustain that the inpainted image in Figure 3(b) is not a “good” solution, it is a
fact that most humans will prefer the image in Figure 2(b) over the one in Figure 3(b) as
an inpainting of the image in Figure 3(a).

This puts into evidence that any inpainting process can entail a high level of subjectivity. That being said, a “good” inpainting method could be defined as one that would most frequently emulate what most humans would do. The rest of the article is strongly aligned with this belief.

1.3 Curvature Driven Diffusion Inpainting

An isophote of an inpainted image represented by \( u \) can be roughly defined as any level line of \( u \) that separates regions of different light intensities. We are looking for an inpainting method that results in an image with isophotes connected inside the occlusion areas, since this will imply that the edges “broken” by the occlusions will be connected. If we observe Figure 3(b), we note that the original isophotes (the edges of the black bar) do not result connected as we would most probably hope. Note also that the inpainted isophotes have corners, meaning that their curvature \( \kappa \) at those points is \( \pm \infty \), in contrast with the zero curvature in the isophotes of the “expected” inpainted image (Figure 2(b)).

This led Chan and Shen [4] to consider the curvature in the CDD diffusion model.

Let us have a closer look at equation (4). If we let \( \hat{D} = \frac{1}{|\nabla u|} \), then (4) reads

\[
\frac{\partial u}{\partial t} = \nabla \cdot \left[ \hat{D} \nabla u \right].
\] (5)

The ideas in [4] consist in appropriately modifying the diffusion coefficient \( \hat{D} \) in order to take curvature into account. More precisely, we take \( \hat{D} = g(|\kappa|) \frac{1}{|\nabla u|} \), where \( g : \mathbb{R}_0^+ \to \mathbb{R}_0^+ \) is an increasing function such that \( g(0) = 0 \) and \( g(\infty) = \infty \). With this choice, diffusion is strong where curvature is large, while it is weak where curvature is small. Since the curvature is \( \pm \infty \) at the corners of the TV inpainted image in Figure 3(b), such an image cannot be the steady state of equation (5) with the modified diffusion coefficient \( \hat{D} \). It turns out that in this case the method strongly favors connectivity on the steady state, as can be seen in Figure 4. Note, however, that since this method tends to produce piecewise constant restorations, it is unsuitable for inpainting occlusions over smooth regions (see Figure 4). The main objective of this article is to develop an inpainting method suitable for images which present occlusions over both edges and smooth regions.

Figure 4: Left: occluded image. Right: CDD restoration.
Regularization methods can be combined by using spatially-varying weighted averages of their penalizers [5] to include some characteristic properties of each one. We shall consider the case of mixed T1 and TV regularization, taking advantage of the fact the the first one tends to produce smooth restorations, while the latter is better for edge preservation. The inpainted image will thus be defined as the minimizer of

$$J(u) = \|Tu - v\|_{L^2}^2 + \lambda_{T1}\|\sqrt{1 - \theta} \nabla u\|_{L^2}^2 + \lambda_{TV}\|\lambda A\nabla u\|_{L^1},$$

(6)

where \(\lambda_{T1} > 0\) and \(\lambda_{TV} > 0\) are regularization parameters, \(A = A(x,y)\) is a space dependent anisotropy matrix field and \(\theta = \theta(x,y) \in [0,1]\) is a spatially-varying function weighting penalizations at each point. Note that \(\theta \equiv 0\), corresponds to pure T1 regularization, while \(\theta \equiv 1\), leads to pure anisotropic TV regularization.

The matrix field \(A : \Omega \to \mathbb{R}^{2 \times 2}\), as defined in [1], introduces anisotropy to the regularization model. It is built upon an \(a\)-\(priori\) estimation of the gradient field, \(\nabla u_p(x,y)\), and has the following properties:

- \(A(x,y)\) is a symmetric positive definite matrix \(\forall (x,y) \in \Omega\)
- If \(\nabla u_p(x,y) = 0\), \(A(x,y) = I\) (the identity matrix).
- If \(\nabla u_p(x,y) \neq 0\), \(A(x,y)\) has eigenvalues \(\sigma_j(x,y)\) and eigenvectors \(v_j(x,y)\), such that
  
  \[v_1(x,y) \parallel \nabla u_p(x,y), \quad \sigma_1(x,y) = h(|\nabla u_p(x,y)|)\]
  \[v_2(x,y) \perp \nabla u_p(x,y), \quad \sigma_2(x,y) = 1\]

- The function \(h\) above is decreasing, with \(0 < h(t) \leq 1\) \(\forall t \in \mathbb{R}^+\).

Given the function \(h\), the matrix field \(A\) can be explicitly written as:

$$A(x,y) = I - (1 - h(|\nabla u_p(x,y)|)) \begin{bmatrix} \nabla u_p(x,y) \\ |\nabla u_p(x,y)| \end{bmatrix}^T \begin{bmatrix} \nabla u_p(x,y) \\ |\nabla u_p(x,y)| \end{bmatrix}. \tag{7}$$

In our case, we took \(h(t) = 1/(1 + (t/\tau)^k)\), where \(\tau, k > 0\) are control parameters that could roughly be referred to as the break point from which we infer the image has an edge and the width of the transition region, respectively.

In [1], the gradient field estimation \(\nabla u_p\) is built by inpainting the components of the gradient using a second order Tikhonov-Phillips method. However, since this method produces smooth restorations, edge preservation results compromised for wide occlusions, as seen in Figure 1(c). To overcome this disadvantage, we shall build the matrix field \(A\) based upon the gradients of an \(a\) \(priori\) CDD inpainting of the image, which favors both edge preservation and object connectivity.
In the same fashion, the space variant weighting function $\theta$, which also requires of a previous estimation of the gradient field, can be built from the CDD-inpainted image $u_p$, that is, $\theta(x, y) = w(|\nabla u_p(x, y)|)$, where $w$ is an appropriately chosen increasing function, with $0 \leq w(t) \leq 1$ $\forall t \in \mathbb{R}^+$. 

The full inpainting process can be then stated as follows:

**Step 1: CDD inpainting.** Perform a CDD inpainting on the occluded image to obtain a first estimation. That is, compute $u_p$ as the steady state of the equation

$$\frac{\partial u}{\partial t} = \nabla \cdot \left[ \frac{g(\kappa)}{|\nabla u|} \nabla u \right].$$

**Step 2: Construction of the anisotropy matrix field.** Use $u_p$ to construct $A$ as

$$A = I - \left(1 - \frac{1}{1 + (|\nabla u_p|/\tau)^k}\right) \left[\frac{\nabla u_p}{|\nabla u_p|}\right] \left[\frac{\nabla u_p}{|\nabla u_p|}\right]^t.$$

**Step 3: Construction of the weighting function.** Use $u_p$ to build $\theta$ as

$$\theta = w(|\nabla u_p|).$$

**Step 4: Final inpainting.** Use the previously computed $A$ and $\theta$ to build the mixed weighted T1-TV anisotropic functional and compute the restored image as

$$\hat{u} = \arg\min \left\{ \|Tu - v\|_2^2 + \lambda_1 \|\sqrt{1 - \theta} \nabla u\|_2^2 + \lambda_{TV} \|\theta A \nabla u\|_{L1} \right\}. \quad (8)$$

Having stated the inpainting process, we shall now proceed to briefly explain how it can be numerically implemented.

### 3 NUMERICAL IMPLEMENTATION

The first step to begin treating the inpainting problem is to perform a discretization over the image. Firstly, we assume that the grayscale image domain is $\Omega = [0, 1] \times [0, 1]$ and we discretize it to obtain an $M \times M$ pixel grid and an $M \times M$ matrix $U$, consisting its entries of the values of the function $u$ at the centerpoints of the pixels. Next, we stack the columns of the matrix $U$ to get a vector $u \in \mathbb{R}^{M^2}$ so that $u_{M(l-1)+m} = U_{m,l}$ $\forall l, m = 1, 2, \ldots, M$. For better understanding we will often identify $u_{M(l-1)+m}$ with $u(x, y)$.

#### 3.1 CDD implementation

Let us define $\xi = \frac{u(|\nabla u|)}{|\nabla u|} \nabla u$, whose divergence we will need to approximate. We do so by computing

$$\nabla \cdot \xi(x, y) \approx \frac{\xi_1(x + h/2, y) - \xi_1(x - h/2, y)}{h} + \frac{\xi_2(x, y + h/2) - \xi_2(x, y - h/2)}{h}, \quad (9)$$
where $h = 1/M$ is the pixel-width.

Now, in order to compute $\xi$ at the midpoints between adjacent pixels, we need to estimate (for instance) $\nabla u(x + h/2, y)$. We do this as follows:

$$
\begin{align*}
&u_x(x + \frac{h}{2}, y) \approx \frac{u(x + h, y) - u(x, y)}{h}, \\
&u_y(x + \frac{h}{2}, y) \approx \frac{1}{2}(u(x + h, y + h) + u(x, y + h)) - \frac{1}{2}(u(x + h, y - h) + u(x, y - h)).
\end{align*}
$$

For points of the form $(x, y + h/2)$, the construction is analogous.

As for the isophote curvature $\kappa$, we compute it explicitly as a function of the gradient of $u$:

$$
\kappa = \nabla \cdot \left[ \frac{\nabla u}{|\nabla u|} \right] = \frac{\partial u_x}{\partial x}|\nabla u| + \frac{\partial u_y}{\partial y}|\nabla u|.
$$

For points of the form $(x + h/2, y)$,

$$
\kappa(x+h/2, y) = \frac{1}{h} \left[ \frac{u_x(x+h,y)}{|\nabla u(x+h,y)|} - \frac{u_x(x,y)}{|\nabla u(x,y)|} \right] + \frac{1}{2h} \left[ \frac{u_y(x+h/2,y+h)}{|\nabla u(x+h/2,y+h)|} - \frac{u_y(x+h/2,y-h)}{|\nabla u(x+h/2,y-h)|} \right].
$$

The computation of $\kappa$ at points of the form $(x, y + h/2)$ is analogous.

We are then ready to state the algorithm for the CDD inpainting process. It is timely to point out here that due to numerical stability issues, instead of using Euler’s method to solve the IVP (as was stated in [4]), we use an Adams-Moulton Adams-Bashford prediction-correction method, which shows better results. Our modified algorithm reads as follows:

**Step 1: Initializing.** Define a compatible initial estimation $u^{(0)}$ and let $n = 0$.

**Step 2: Updating.** Compute $u^{(n+1)}_m$ as follows:

$$
\begin{align*}
u^{(n+1)}_m &= u^{(n)}_m + \frac{\Delta t}{12} \left[ 23 \nabla \cdot \left[ \frac{g(|\kappa^{(n)}_m|)}{|\nabla u^{(n)}_m|} \nabla u^{(n)}_m \right] - 16 \nabla \cdot \left[ \frac{g(|\kappa^{(n-1)}_m|)}{|\nabla u^{(n-1)}_m|} \nabla u^{(n-1)}_m \right] + 5 \nabla \cdot \left[ \frac{g(|\kappa^{(n-2)}_m|)}{|\nabla u^{(n-2)}_m|} \nabla u^{(n-2)}_m \right] \right], \\
u^{(n+1)}_m &= u^{(n)}_m + \frac{\Delta t}{12} \left[ 5 \nabla \cdot \left[ \frac{g(|\kappa^{(n+1)}_m|)}{|\nabla u^{(n+1)}_m|} \nabla u^{(n+1)}_m \right] + 8 \nabla \cdot \left[ \frac{g(|\kappa^{(n)}_m|)}{|\nabla u^{(n)}_m|} \nabla u^{(n)}_m \right] - \nabla \cdot \left[ \frac{g(|\kappa^{(n-1)}_m|)}{|\nabla u^{(n-1)}_m|} \nabla u^{(n-1)}_m \right] \right].
\end{align*}
$$

where $\kappa^{(n)}$ and $\nabla u^{(n)}$ are computed from $u^{(n)}$ as previously stated. For the first two steps, we use a second order Runge-Kutta method.

**Step 3: Stopping.** If an appropriate stopping criterion (defined upon the decay on the total curvature of the image) is reached, the process stops and the inpainted image is defined as $\hat{u} = u^{(n+1)}$. Else, $n$ is increased ($n = n + 1$) and the algorithm iterates from Step 2.

Next, we will show how the mixed weighted anisotropic T1-TV regularization can be numerically implemented.
3.2 Mixed anisotropic T1-TV implementation

To find the minimizer of the T1-TV inpainting functional (more details can be found in [7]), let us consider the discretized version of functional (6)

\[ J(u) = \| Tu - v \|^2 + \lambda T_1 \sum_{m \in \mathcal{M}} \sqrt{1 - \theta_m} \left[ \left( \frac{u_m + M - u_{m+1}}{1/M} \right)^2 + \left( \frac{u_{m-1} - u_m}{1/M} \right)^2 \right] + \lambda_{\text{TV}} \sum_{m \in \mathcal{M}} \theta_m \left\| A_m \left( \frac{M(u_m - u_{m+1})}{M(u_m - u_{m-M})} \right) \right\|_1, \]

where \( T \in \mathbb{R}^{M^2 \times M^2} \) is the diagonal matrix associated to the occlusion operator \( T \), \( \theta_m \) and \( A_m \) are the values of the weighting function and the anisotropy matrix field on the \( m^{th} \) pixel, and \( \mathcal{M} \) is the set of the interior pixels, on which the gradient is estimated. Naturally, the restored image will be approximated by the minimizer of this functional. In order to find such a minimizer, we will approximate \( J \) by a differentiable functional and make use of a duality relation to approach its minimizer iteratively. Let us begin by noting that the \( m^{th} \) term of the discretized anisotropy penalizer on (10) can be written as

\[
\left\| A_m \left( \frac{M(u_m - u_{m+1})}{M(u_m - u_{m-M})} \right) \right\|_1 = |M(a_{1,1}^m(u_m - u_{m-M}) + a_{1,2}^m(u_m - u_{m+1}))| + |M(a_{2,1}^m(u_m - u_{m-M}) + a_{2,2}^m(u_m - u_{m+1}))|,
\]

where \( a_{i,j}^m \) is the \((i, j)\) component of \( A_m \). We would like now to minimize (10) simply by solving its first order necessary condition, but since the absolute value in (11) is non-differentiable, we shall approximate it by a differentiable function \( |t| \approx \phi(t) \doteq \sqrt{t^2 + \eta^2} - \eta \) for \( \eta > 0 \) sufficiently small. With this choice of \( \phi \), there exists a function \( \psi \) that satisfies

\[
\phi(t) = \inf_{s > 0} (st^2 + \psi(s)), \tag{12}
\]

\[
\psi(s) = \sup_{t \in \mathbb{R}} (\phi(t) - st^2).
\]

By these means, (11) can be approximated as follows

\[
\left\| A_m \left( \frac{M(u_m - u_{m+1})}{M(u_m - u_{m-M})} \right) \right\|_1 \approx \inf_{s \in \mathbb{R}^{M^2}} \left( s_m \left( \frac{a_{1,1}^m(u_m - u_{m-M}) + a_{1,2}^m(u_m - u_{m+1})}{1/M} \right)^2 + \psi(s_m) \right) + \inf_{q \in \mathbb{R}^{M^2}} \left( q_m \left( \frac{a_{2,1}^m(u_m - u_{m-M}) + a_{2,2}^m(u_m - u_{m+1})}{1/M} \right)^2 + \psi(q_m) \right) \tag{13}
\]

Define now the \( M^2 \)-by-\( M^2 \) diagonal matrices \( A^{i,j} \) for \( i, j = 1, 2 \), such that \( A_{m,m}^{i,j} = a_{i,j}^m \) if \( m \in \mathcal{M} \) and \( A_{m,m}^{i,j} = 0 \) otherwise. Let \( L_x \) and \( L_y \) be the \( M^2 \)-by-\( M^2 \) approximating matrices.
for the components of the gradient, and let \( R_1 \) and \( R_2 \) be the \( M^2 \)-by-\( M^2 \) matrices defined by \( R_1 = A^{1,1}L_x + A^{1,2}L_y \) and \( R_2 = A^{2,1}L_x + A^{2,2}L_y \). Finally, we let \( I \) be the \( M^2 \)-by-\( M^2 \) identity matrix, and define the functional
\[
K_{\theta,\phi}(u, s, q) = \|Tu - v\|^2 + \frac{\lambda_{T1}}{M^2} u' L_x(I - \Theta)L_x + L_y(I - \Theta)L_y)u + \frac{\lambda_{TV}}{M} u'^R_1 \Theta S R_1 u
+ \frac{\lambda_{TV}}{M} \sum \theta_m \psi(s_m) + \theta_m \psi(q_m).
\]
(14)

It can be shown ([7]) that
\[
\inf_{s,q \in \mathbb{R}^{M^2}} K_{\theta,\phi}(u, s, q) = J_\phi(u) \approx J(u),
\]
(15)

where \( J_\phi \) is the functional \( J \) with the absolute value replaced by \( \phi \). Now the problem is equivalent to minimizing \( K \) with respect to \( u, s \) and \( q \) simultaneously. We begin by noting that the first order necessary condition over \( K \) with respect to \( u \) can be written as
\[
\left( T^T + \frac{\lambda_{T1}}{M^2} (L_x(I - \Theta)L_x + L_y(I - \Theta)L_y) + \frac{\lambda_{TV}}{M^2} u'^R_1 \Theta S R_1 + \frac{\lambda_{TV}}{M^2} u'^R_2 \Theta S R_2 \right) u = T^Tv.
\]
(16)

In order to minimize \( K_{\theta,\phi}(u, s, q) \) with respect to \( s \) and \( q \), we define
\[
b_m = \arg\min_{s_m \in \mathbb{R}_+} \left( s_m M^2 \left( a_{1,1}^m (u_m - u_{m-M}) + a_{1,2}^m (u_m - u_{m+1}) \right)^2 + \psi(s_m) \right),
\]
and
\[
c_m = \arg\min_{q_m \in \mathbb{R}_+} \left( q_m M^2 \left( a_{2,1}^m (u_m - u_{m-M}) + a_{2,2}^m (u_m - u_{m+1}) \right)^2 + \psi(q_m) \right).
\]

We now resort to (12) to deduce that the elements \( b_m \) and \( c_m \) must satisfy
\[
b_m = \frac{\phi'(M \left( a_{1,1}^m (u_m - u_{m-M}) + a_{1,2}^m (u_m - u_{m+1}) \right))}{2M \left( a_{1,1}^m (u_m - u_{m-M}) + a_{1,2}^m (u_m - u_{m+1}) \right)}
\]
(17)
and
\[
c_m = \frac{\phi'(M \left( a_{2,1}^m (u_m - u_{m-M}) + a_{2,2}^m (u_m - u_{m+1}) \right))}{2M \left( a_{2,1}^m (u_m - u_{m-M}) + a_{2,2}^m (u_m - u_{m+1}) \right)}.
\]
(18)

Based upon all of the above, we state a cyclic iterative algorithm for the T1-BV image inpainting as follows:

**Step 1: Initializing.** Set \( j = 0 \), and initialize \( u^j = u^0 \).

**Step 2: Counting.** Let \( j = j + 1 \).
Step 3: Updating $b^j_m$ and $c^j_m$. Update $b^j_m$ and $c^j_m$ using equations (17) and (18).

$$b^j_m = \phi'(M(a_{1,1}^m(u^j_m - u^{j-1}_{m-M}) + a_{1,2}^m(u^{j-1}_m - u^{j-1}_{m+1})))$$
$$2M(a_{1,1}^m(u^{j-1}_m - u^{j-1}_{m-M}) + a_{1,2}^m(u^{j-1}_m - u^{j-1}_{m+1}))$$

$$c^j_m = \phi'(M(a_{2,1}^m(u^{j-1}_m - u^{j-1}_{m-M}) + a_{2,2}^m(u^{j-1}_m - u^{j-1}_{m+1})))$$
$$2M(a_{2,1}^m(u^{j-1}_m - u^{j-1}_{m-M}) + a_{2,2}^m(u^{j-1}_m - u^{j-1}_{m+1}))$$

Step 4: Updating $u$. Update $u^j$ by solving the linear system

$$\left(T^4 + \frac{\lambda TV}{M^2}(L_x(I - \Theta) + L_y(I - \Theta)) + \lambda TV R_1^j B^j R_1 + \lambda TV R_2^j C^j R_2\right) u^j = T^4 v$$

where $B^j$ and $C^j$ are the $M^2$-by-$M^2$ diagonal matrices with elements $b^j_m$ and $c^j_m$ for $m \in \mathcal{M}$ and 0 otherwise.

Step 5 - Convergence: if a previously defined convergency criterion is satisfied, the algorithm ends and our restored signal is defined as $u^j$. Else, the algorithm repeats from step 2.

Up next, we will show some examples of the performance of the inpainting method.

4 NUMERICAL RESULTS

We begin by comparing our new approach with the T1 and the CDD methods on the previously used test image. Figure 5 depicts the grayscale image occluded over both smooth and piecewise constant areas and the results of the three different inpainting processes.

(a) (b) (c) (d)

Figure 5: Occluded image (a); T1 inpainting (b); CDD inpainting (c); mixed T1-TV inpainting (d).

The mixed T1-TV inpainting clearly outperforms the T1 method in terms of edge preservation, while it also works better than the CDD inpainting for the smooth area of the image. Although this is a somewhat subjective analysis, its conclusions can be supported by appropriately analyzing the ISNR values.
We will now show an example of the performance of our new method on a “real” image. Here, the inpainting process was performed separately over the red, green and blue layers of a $150 \times 150$ color image. In this case, a $\%1$ Gaussian white noise was added to the occluded image. Figure 6 shows the occluded noisy image, along with the T1, CDD and mixed T1-TV inpainting images for comparison purposes. The optimal regularization parameters obtained with Morozov’s Principle are also depicted.

<table>
<thead>
<tr>
<th>Layer</th>
<th>$\lambda_{T1}$ for T1</th>
<th>$\lambda_{T1}$ for T1-TV</th>
<th>$\lambda_{TV}$ for T1-TV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red</td>
<td>$2.98 \times 10^{-6}$</td>
<td>$5.04 \times 10^{-6}$</td>
<td>$2.58 \times 10^{-4}$</td>
</tr>
<tr>
<td>Green</td>
<td>$2.22 \times 10^{-6}$</td>
<td>$3.86 \times 10^{-6}$</td>
<td>$3.15 \times 10^{-4}$</td>
</tr>
<tr>
<td>Blue</td>
<td>$2.94 \times 10^{-6}$</td>
<td>$5.88 \times 10^{-6}$</td>
<td>$2.00 \times 10^{-4}$</td>
</tr>
</tbody>
</table>

Figure 6: Occluded noisy image (a); T1 inpainting (b); CDD inpainting (c); mixed T1-TV inpainting (d).

Once again we observe that the mixed T1-TV inpainting outperforms both the T1 and CDD methods. Since we cannot compute the ISNR of the images as there is no information about the occluded region, Table 1 shows the norm of the differences between the restorations and the original images as a formal comparison tool. While the original image in Figure 6 results rather obvious to infer, the “most-humans” choice was consider as the real test image.
Table 1: Performance comparisons.

<table>
<thead>
<tr>
<th></th>
<th>Gray Img</th>
<th></th>
<th>Color Img</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T1</td>
<td>CDD</td>
<td>T1-TV</td>
<td>T1</td>
</tr>
<tr>
<td>$|\hat{u} - u_0|$</td>
<td>8.89</td>
<td>1.19</td>
<td>0.34</td>
<td>8.79</td>
</tr>
</tbody>
</table>

5 CONCLUSIONS

In this article, we developed a two-step method for image inpainting consisting of a CDD process followed by a mixed anisotropic T1-TV regularization method. The first step is used to include structural information into the second one. Applications of this new combined method showed that its performance is significantly better than the other single methods in the sense that it better complies with the “human-eye inpainting paradigm”. Improvements on the ideas presented in this article are currently under development.

6 ACKNOWLEDGEMENTS

This work was supported in part by Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET, through PIP 2014-2016 Nro. 11220130100216-CO and by the Air Force Office of Scientific Research, AFOSR/SOARD, through Grant FA9550-14-1-0130.

REFERENCES


COMPREHENSIVE COMPARISON OF PERFORMANCE OF ELITE-INDUCED EVOLUTIONARY ALGORITHM FOR MULTI-OBJECTIVE ANALYSIS

Y. SHIMIZU¹ AND R. NAKAMURA²

¹ Toyohashi University of Technology
1-1 Hibarigaoka, Tenpaku-cho, Toyohashi, Aichi, JAPAN
shimizu@me.tut.ac.jp, http://ise.me.tut.ac.jp

² Toyohashi University of Technology
nakamura@ise.me.tut.ac.jp

Key words: Multi-objective Analysis, Elite Induction, Aggregating Method, Multi-objective Evolutionary Algorithm, Post-optimal Analysis.

Abstract. To cope with manifold and complicated decision making in today society, multi-objective optimization has gained greater and greater concerns recently. As a family of such studies, we engaged in multi-objective analysis and proposed a new idea termed elite-induced evolutionary algorithm. The proposed method is a hybrid method comprised of a certain conventional aggregating method and recent one referring to multi-objective evolutionary method. Through a simple but complemented combination of them, it is able to derive a precise Pareto front efficiently while even controlling a set of the Pareto optimal solutions under the preference of decision maker. In previous study, we showed these distinguished features only for PSA as the evolutionary method. To ascertain such prospect, in this paper, we will evaluate the performance with a comprehensive point of view. For this purpose, besides PSA, we have deployed the idea over several methods popularly known, i.e. MOPSO, MODE, NSGA II and SPEA II. Moreover, we provided a few indices to quantitatively and commonly evaluate the performance among these different methods. Finally, numerical experiments have been carried out based on a few simple test problems and the constrained optimization problems.

1 INTRODUCTION

To cope with manifold and complicated decision making in today society, multi-objective optimization (MOP) has been in higher demand recently. Notably, MOP becomes especially amenable to optimization problems involving incommensurable objectives that conflict or compete with each other. In fact, it is viewed as a promising decision aid from comprehensive point of view in global and competitive manufacturing tasks. As a family of such MOP studies, we engaged in multi-objective analysis (MOA) and proposed a new idea termed elite-induced evolutionary algorithm for MOA (EIEA/MAO) [1,2]. Here, we define MOP as a subjective synthesis to obtain a unique preferentially optimal solution or best compromise solution of decision maker (DM) [3]. Meanwhile MOA is viewed as an objective
analysis to reveal a trade-off relation among the conflicting objectives or simply to draw the Pareto front that is a set of infinite Pareto optimal solutions [4]. Since the Pareto optimal solution represents a rational norm in MOP, a set of such solutions can provide DM significantly useful information to accomplish MOP as a final goal of the decision making.

In our previous study, we showed the effectiveness of our method by concerning only with PSA (Pareto Simulated Annealing) [5] as an evolutionary algorithm. To ascertain some favorable properties revealed thereat, this study aims at totally evaluating the effectiveness of EIEA/MOA. For this purpose, we have deployed the idea over additional MOEA methods popularly applied to various problem-solving, i.e., MOPSO [6], MODE [7], NSGA II [8] and SPEA II [9] besides PSA. Moreover, we provided a few indices to quantitatively evaluate the performance among these different methods commonly. To discuss such aspect, numerical experiments have been carried out based on a simple test problem first. Then, we turned our attention to more practically important topic in engineering, i.e., constrained optimization problems. Through these, we try to reveal some suggestive features of EIEA/MOA and finally confirm its effectiveness totally.

The rest of this chapter is organized as follows. In Section 2, some general statements are described associated with the present concern. Section 3 deploys the idea in terms of the elements necessary for discussion. In Section 4, numerical experiments provide manifold comparisons over the variants of method. A conclusion is given in Section 5.

2 GENERAL STATEMENTS ON MULTI-OBJECTIVE ANALYSIS ASSOCIATED WITH PRESENT CONCERN

The generic MOP can be described as follow.

\[
\min \left\{ f_1(x), f_2(x), \ldots, f_N(x) \right\} \quad \text{subject to} \quad x \in X,
\]

where \( x \) denotes an \( n \)-dimensional decision variable vector, \( X \) an admissible region and \( f_i \), \((i=1,\ldots, N)\) objective functions some of which are incommensurable and conflict with one another.

To obtain a unique preferentially optimal solution or to solve the above MOP, we need some subjective judgments on preference of DM. This is the most distinguished difference compared with the usual or single-objective optimization and MOA as well. They are completely free from such subjectivity. At the stage of MOP, MOA can probably provide the most straightforward and useful information. Regardless of such fundamental understanding, the extension of evolutionary algorithms (EA) has caused greater interests and various multi-objective evolutionary algorithms (MOEA) have been developed. As pointed out already, these methods are viewed as a MOA that tries to reveal a certain tradeoff relation among the conflicting objectives. Hence, when working with this problem, we have potentially the following three major goals to achieve [10].

1. Maximize the number of elements of the Pareto optimal solution set (POSS)
2. Minimize the difference of the Pareto front produced by the method from the true one
3. Make the distribution of the elements of POSS as smooth and uniform as possible

In the below, we briefly introduce some methods for MOA applied in our idea of EIEA since readers can easily find the detail of those methods from many text books. Anyone
belonging to the aggregating method is a conventional method that tries to transform the original multi-objective problem into a single-objective problem by unifying the objective functions in an appropriate manner. For example, \( \varepsilon \)-constraint method, weighting method and weighted Min-Max method are popularly known. Then, the resulting problem is solved recursively by a certain mathematical programming method while changing certain parameters like weighting coefficients. For further development, we conveniently applied the \( \varepsilon \)-constraint method described below.

\[
\begin{align*}
\text{min } f_p(x) \quad \text{subject to} \quad & \begin{cases} 
x \in X \\
f_i(x) \leq f_i^* + \varepsilon_i, \quad (i = 1, 2, \ldots, N, i \neq p).
\end{cases}
\end{align*}
\]

where \( f_p(x), f_i^* \) and \( \varepsilon_i \) denote the principal objective function, minimum value of \( f_i(x) \) and its relaxed amount, respectively. We can draw an accurate POSS by solving the above problem repeatedly with a set of relaxed amounts of \( \varepsilon_i \). So it is somewhat possible to manipulate the distribution of solutions mentioned in the above.

On the other hand, we note the following five methods of MOEA that are popularly known and have been widely used.

1. Pareto Simulated Annealing (PSA): Though the general procedure of PSA is similar to that of the single-objective SA, the algorithm is devised to accelerate a tendency for approaching the Pareto front as well as a dispersion of the solution over the whole front. To work with this, updating weighting coefficients among the objectives becomes a key point besides the probability of acceptance of the neighbor solution.

2. Multi-Objective Particle Swarm Optimization (MOPSO): The single-objective PSO is a technique after the behavior of bird flocking or fish schooling (swarm). Members of swarm communicate with each other and adjust their own positions and velocities based on the information regarding good positions both of their own and the swarm as well. When applying such ideas to multi-objective problems, a few modifications must be made for the original algorithm. First, we should note on a set of solutions comprising the Pareto front instead of one global best solution. To work with this, a set of leaders which are non-dominated solutions found at each iteration are stored in an external archive. Then, some sort of quality measure is calculated for all the leaders to select one leader for each particle of the swarm. At each generation, a leader is selected for each particle and the flight is performed.

3. Multi-Objective Differential Evolution (MODE): The original differential evolution (DE) is viewed as a real variable variant of genetic algorithm. This method is evolved from such DE by combining Pareto dominance principles and using elitism in its evolution. The main algorithm consists of initialization of population, evaluation, Pareto dominance selection, performing DE operations and repeating the search on population to reach the Pareto optimal solutions. One of the crucial points of this method is that an external archive is used to store the non-dominated solutions found so far over the generation.

4. Non-dominated Sorting Genetic Algorithm II (NSGA II): This method uses the idea of elitism that can avoid both deleting the superior solutions found previously and crowding to maintain the diversity of solutions. In this method, non-dominated sorting is carried out for all members of the parents and offspring populations together with a crowding distance which is an estimate of the density of solutions neighboring a particular solution. By virtue of these operators, NSGA II is considerably faster than its predecessor NSGA and gives very good
results for many problems.

5. Strength Pareto Evolutionary Algorithm II (SPEA II): SPEA II is an improved version of its predecessor SPEA, which incorporates a fine-grained fitness assignment strategy, a density estimation technique, and an enhanced archive truncation method. The main differences compared with SPEA are an improved fitness assignment scheme, a nearest neighbor density estimation technique and a new archive truncation method that will guarantee the preservation of boundary solutions on Pareto front.

3 ELITE INDUCED MOEA

3.1 Basic idea

The principle behind the general idea of EIEA/MOA is just simple and straightforward from the original MOEA, i.e., a plain combination of a traditional aggregating method and a MOEA. However, it can make the best use of the advantages inherited by both methods while virtually eliminating shortcomings of each method (Refer to Table 1). The original MOEA algorithm serves as a main driving algorithm in EIEA/MOA. One of the major differences is that instead of using all randomly generated initial solutions, it introduces some number of the elite solutions that are obtained from the aggregating method. Accordingly, the algorithm is composed of two parts:

1. Generation of elite solutions using the aggregate method under a certain allocation rule mentioned below.
2. Application of MOEA by incorporating the elite solutions into a set of random initial solutions.

General scheme of EIEA/MOA is demonstrated in Fig. 1 for two-dimensional objective space. As depicted there, various locations of elites are predetermined by the allocation rules guiding dominated solutions to the elites. They are referred to normal, North-West/South-East corners, and Pareto corner distribution, respectively. We expect each elite solution set will induce the Pareto front at the direction toward its preexisting region. By adjusting the number of elites and the allocation rule appropriately, DM is able to manipulate a location of final solutions so that the solutions would lie on a specific region on Pareto front. This makes EIEA/MOA unique since it introduces something subjective in the objective task of MOA and provides more degree of freedom and flexibility to concern with the diversified value system of DM. Due to the existence of the elites, selection pressure that might contribute to the accuracy and convergence speed is always kept at high level. This makes the algorithm powerful and computation load smaller. Moreover, the induction ability that will easily guide the solution set to DM’s preferable region is a unique function not concerned at all previously. In a summary, the major concerns that will affect on its performance refer to the number of elite solutions or its rate to the whole size and their locations. The option of selecting the most

![Figure 1: The mechanism of EIEA/MOA and elite allocation rules](image-url)
relevant MOEA is also notable aspect of the proposed approach.

<table>
<thead>
<tr>
<th>Method</th>
<th>Advantage</th>
<th>Disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggregating</td>
<td>Accurate Pareto front can be obtained by powerful NLP software.</td>
<td>Dispersion adjustment of the Pareto front is necessary and a bit tedious.</td>
</tr>
</tbody>
</table>
| MOEA            | No adjustment in dispersion of the Pareto front is required | • Poor accuracy of the Pareto front
|                 |                                                 | • Slow convergence                                 |

3.2 Significance of induction

Though we expect to have at most several candidate solutions available for the final decision making, every MOEA will derive a set of infinite solutions. This may refer to a fact that MOEA is out of concern about any preference imbedded in DM, and earnestly pays its effort on realizing a uniform distribution over the entire Pareto front. However, even in MOA, we had better address the DM’s preference more elaborately. In reference to such opinion, a discussion of diversification may be meaningful just on the front where any more preferable points will not exist if they are off the front. However, it does not make sense if this is not the case [11]. The elite induction technique can readily manage this prior discrimination that becomes helpful in the next MOP phase. Moreover, such induction ability is available for a handy post-optimal analysis of MOP when the preferentially optimal solution is assigned as the single elite. In fact, we can easily obtain various solutions around it by adjusting the archive size and the niche technique properly.

3.3 Performance Metrics (PMs)

As mentioned in Section 2, performances of MOA are to be usually evaluated in terms of the three indexes, i.e., number of non-dominated solutions in the obtained set, distance to the true Pareto front and the uniform distribution. The generational distance [12] is a metric of calculating the average distance between the solutions found and those on the Pareto front. If this value is zero, it means that all the solutions found are on the Pareto front. Spacing metric [13] judges how well the solutions are distributed on the Pareto front. If this metric takes zero, all solutions are equidistantly spaced on the Pareto front. Due to some reasons, however, these conventional metrics are not directly amenable for the present concerns. That is, it is not always possible to have a true Pareto front (in most real-world applications); an irregular but stretched distribution may be more preferable to a uniform but narrow one (Both uniformity and stretch should be evaluated); we can neglect the number of solutions if we view MOA as a prior stage of MOP (Rather small number is more preferable for MOP). From these reasons, in this study, we provide the following four metrics.

3.3.1 Spacing; $SP$

To evaluate the stretch of Pareto front beside the uniformity (Refer to Fig.2), we augment the $N$ supplemental solutions termed extreme points by solving $N$ single-objective optimization problems such as “$\min f(x)$ s.t. $x \in X (i=1,...,N)$”, in turn. Here, the extreme point is given by evaluating the objective values through the respective optimal solution $x^*_i$, i.e.,
\[ f^1 = f^{m+1} = (f_1(x_1^*), f_2(x_2^*), ..., f_N(x_N^*))^T, \]
\[ \ldots \]
\[ f^N = f^{m+N} = (f_1(x_1^*), f_2(x_2^*), ..., f_N(x_N^*))^T. \]

Now, SP is defined as
\[ SP = \frac{1}{m - 1 + N} \sum_{i=1}^{m+N} (\bar{d} - d'_i)^2 \]

where \( m \) is the number of non-dominated solutions and \( \bar{d} \) the mean of all \( d'_i \). The smaller this value is, the more uniformed distribution could be gained over the extended common space.

### 3.3.2 Piercing index; PI

Piercing index is defined by
\[ PI = \frac{\sum_{j=1}^{N} \sum_{k=1}^{m} |f_{j}^{*}(x) - f_{j}^{k}(x)|^{N}}{m} \]

where \( f_{j}^{*} \) denotes a utopia point generated by collecting the optimal objective values from \( N \) single-objective optimizations, i.e., \( f_{j}^{*} = (f_1(x_1^*), f_2(x_2^*), ..., f_N(x_N^*))^T \). \( PI \) represents the average generalized distance between each solution and the utopia point (Refer to Fig.3). Hence, the smaller \( PI \) is, the more accurate Pareto front is obtained.

### 3.3.3 Average rank; AR

Rank is popularly used in the variants of multi-objective genetic algorithm and defined by \( 1 + n_{\text{NDS}} \) [14]. Here \( n_{\text{NDS}} \) denotes the number of dominated solutions for the target solution (See Fig.4). To evaluate a quality of the produced set among the different methods commonly, we introduce the average rank as follows. First accumulate the found solutions from each method into a set. Then, evaluate the rank of every solution of each method in the accumulated set and calculate its average rank. When every solution found by a certain method does not dominated by anyone, this takes the best value, i.e., \( AR = 1.0 \). On the other hand, the members of the set are composed of from \( m \)-worst solutions to the worst, it become \( mM-0.5(m-1) \) if the number of the methods is \( M \) and the found solutions are all \( m \) for them.

### 3.3.4 Induction rate; IR

Induction by the elite is the ability to draw the solutions closely to the elite itself. To measure this strength, we invented the metric termed induction rate \( IR \) as follows.

[Figure 2: 2-dimensional scheme of spacing; solid lines are augmented parts to take the stretch into account]

[Figure 3: 2-dimensional scheme of piercing index]

[Figure 4: 2-dimensional scheme to decide rank]
\[
IR = 1 - \frac{\left(\frac{\sum_{i=1}^{m} l_i}{m}\right)/m}{\sum_{j=1}^{N} \left| f_j^{\text{up}} - f_j^{\text{elite}} \right|^{1/N}}
\]

where \(l_i\) denotes the distance between the found solutions \(i\) and the elite. In the second term of the right hand side of the above equation, this average value is normalized by the generalized distance between the utopia and the elite (Refer to Fig.5). The greater this value is, the stronger induction could be gained. When \(IR=1.0\), all solutions correspond to the elite.

4 NUMERICAL EXPERIMENTS

4.1 Comparison through a simple test problem

In terms of the following test problem [15], we compared the performance on the above metrics regarding the original version and the elite-induced one (E-version) among the methods. The Pareto front of this problem becomes convex when \(k=0.5\) while concave when \(k=2\) and is possible to calculate analytically. In this problem with \(n=10\), population size and numbers of generation are set at 30 and 500, respectively all over the methods and versions. In E-version, five elites are introduced and located between 0.1 and 0.9 with 0.2 intervals along with \(f_1\) axis. The other tuning parameters are set appropriately and their major values are summarized in Table 2. On the other hand, the present extreme points are given as \(f^*=(0.0, 1.0)^T\) and \(f^*==(1.0, 0.0)^T\) and utopia is \(f^{\text{up}}=(0.0, 0.0)^T\).

\[
\begin{align*}
\text{Min} \{f_1(x) = x_1, f_2(x) = g \times h\} \\
g &= 1 + 10 \sum_{i=2}^{n} x_i \\
h &= \begin{cases} 
1 - \left(\frac{f_1}{g}\right)^k, & \text{if } f_1 \leq g \\
0, & \text{otherwise.}
\end{cases} \\
x_i \in [0,1] \quad i = 1, \ldots, n, \quad k = \{0.5, 2.0\}.
\end{align*}
\]

Analytic Pareto front : \(f_1(x) = x_1, f_2(x) = 1 - x_1^k\)

First, we compared the three metrics and summarize those results in Tables 3 and 4 for \(k=0.5\) and \(k=2\), respectively. Thereat, the values in the parenthesis denote the result of the original version of each method. When \(k=0.5\) (convex front), every E-version overwhelmingly outperforms the respective original one in all metrics. On the other hand, when \(k=2\) (concave front), the advantages of E-versions are somewhat limited and we have a bit inferior cases in three cases underlined. At a glance, E-version of MOPSO performs best in the case of convex front while its advantages greatly fall down for the concave front. Nevertheless E-version of MOPSO and PSA can significantly improve the performance from

<table>
<thead>
<tr>
<th>Table 2 Parameter setting for the selection</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Selection</strong></td>
</tr>
<tr>
<td>PSA</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>MOPSO</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>MODE</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>NSGAII</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>SPEAII</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

\textbf{Figure 5:} 2-dimensional scheme of induction rate
their respective original version. On the other hand, every E-version of MODE, NSGA II and SPEA II stably behaves well regardless of the shapes of the front and the versions.

Table 3 Comparison of metrics; Elite-induced (Original) versions and MOEA methods when $k=0.5$

<table>
<thead>
<tr>
<th>Metric</th>
<th>PSA</th>
<th>MOPSO</th>
<th>MODE</th>
<th>NSGA II</th>
<th>SPEA II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spacing</td>
<td>0.062 (0.74)</td>
<td><strong>0.023</strong> (1.00)</td>
<td>0.025 (2.53)</td>
<td>0.057 (0.25)</td>
<td>0.062 (0.34)</td>
</tr>
<tr>
<td>Piercing</td>
<td>0.78 (0.99)</td>
<td><strong>0.59</strong> (3.20)</td>
<td>0.68 (1.99)</td>
<td>0.72 (0.74)</td>
<td>0.78 (0.99)</td>
</tr>
<tr>
<td>Average rank</td>
<td>3.9 (31.0)</td>
<td><strong>2.3</strong> (132.1)</td>
<td>2.4 (20.2)</td>
<td>2.6 (3.3)</td>
<td>2.5 (3.2)</td>
</tr>
</tbody>
</table>

Table 4 Comparison of metrics; Elite-induced (Original) versions and MOEA methods when $k=2$

<table>
<thead>
<tr>
<th>Metric</th>
<th>PSA</th>
<th>MOPSO</th>
<th>MODE</th>
<th>NSGA II</th>
<th>SPEA II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spacing</td>
<td>0.051 (0.16)</td>
<td>0.329 (0.668)</td>
<td><strong>0.027</strong> (0.084)</td>
<td>0.056 (0.052)</td>
<td>0.059 (0.050)</td>
</tr>
<tr>
<td>Piercing</td>
<td>0.96 (0.98)</td>
<td>1.33 (2.93)</td>
<td><strong>0.93</strong> (0.96)</td>
<td>0.936 (0.923)</td>
<td><strong>0.925</strong> (0.927)</td>
</tr>
<tr>
<td>Average rank</td>
<td>3.4 (5.1)</td>
<td>44.2 (173.2)</td>
<td>2.6 (4.3)</td>
<td><strong>2.3</strong> (2.6)</td>
<td>2.5 (3.2)</td>
</tr>
</tbody>
</table>

In the next, we evaluated the induction ability of E-version of each method. For this purpose, we put elite at three specific locations on the Pareto front, i.e., North-West, Pareto corner and South-East. We summarize the results in Tables 5 and 6 for $k=0.5$ and $k=2$, respectively. From these, we can say MODE globally exhibits the most strong induction ability among the methods and finally it can behave as the most amenable E-version if we add the above discussion in terms of the three metrics. On the other hand, NSGA II and SPEA II perform well to drive the Pareto front even in the original version as it is.

Table 5 Summary of induction metric regarding pre-location of elite when $k=0.5$

<table>
<thead>
<tr>
<th>Pre-location</th>
<th>E-PSA</th>
<th>E-MOPSO</th>
<th>E-MODE</th>
<th>E-NSGA II</th>
<th>E-SPEA II</th>
</tr>
</thead>
<tbody>
<tr>
<td>North-West</td>
<td>0.900</td>
<td><strong>0.965</strong></td>
<td>0.866</td>
<td>0.619</td>
<td>0.640</td>
</tr>
<tr>
<td>Pareto corner</td>
<td>0.857</td>
<td><strong>0.967</strong></td>
<td>0.921</td>
<td>0.716</td>
<td>0.724</td>
</tr>
<tr>
<td>South-East</td>
<td><strong>0.961</strong></td>
<td>0.959</td>
<td>0.896</td>
<td>0.557</td>
<td>0.555</td>
</tr>
<tr>
<td>Average</td>
<td>0.906</td>
<td><strong>0.964</strong></td>
<td>0.894</td>
<td>0.631</td>
<td>0.640</td>
</tr>
</tbody>
</table>

Table 6 Summary of induction metric regarding pre-location of elite when $k=2$

<table>
<thead>
<tr>
<th>Pre-location</th>
<th>E-PSA</th>
<th>E-MOPSO</th>
<th>E-MODE</th>
<th>E-NSGA II</th>
<th>E-SPEA II</th>
</tr>
</thead>
<tbody>
<tr>
<td>North-West</td>
<td>0.754</td>
<td>0.347</td>
<td><strong>0.916</strong></td>
<td>0.547</td>
<td>0.542</td>
</tr>
<tr>
<td>Pareto corner</td>
<td>0.793</td>
<td>0.329</td>
<td><strong>0.913</strong></td>
<td>0.693</td>
<td>0.723</td>
</tr>
<tr>
<td>South-East</td>
<td>0.619</td>
<td>0.360</td>
<td><strong>0.906</strong></td>
<td>0.624</td>
<td>0.621</td>
</tr>
<tr>
<td>Average</td>
<td>0.722</td>
<td>0.345</td>
<td><strong>0.912</strong></td>
<td>0.621</td>
<td>0.629</td>
</tr>
</tbody>
</table>

4.2 Comparison through constrained problems

Here, taking the following four constrained problems [16,17], we compared the distribution of Pareto front between E-version and the original one.

1. **TNK problem** is a popular test problem and described as follows.
Min \left\{ f_1(x) = x_1, \ f_2(x) = x_2 \right\}
subject to
\[ g_1(x) = x_1^2 + x_2^2 - 1 - 0.1\cos(16\tan^{-1}(x_1 / x_2)) \geq 0 \]
\[ g_2(x) = -(x_1 - 0.5)^2 - (x_2 - 0.5)^2 + 0.5 \geq 0 \]
\[ 0 \leq x_1, x_2 \leq \pi \]

The Pareto front of this problem is known to be non-convex and discontinuous. The location of the elites are decided from “\( \min f_5(x) \ \text{s.t.} \ f_1(x) \leq 0.1 \text{or} \ f_1(x) \leq 0.5 \text{ or} \ f_1(x) \leq 1.0 \)”.

2. **2-bar truss problem** is an engineering design problem formulated below (Fig.6). This tries to minimize the volume of material \( f_1 \) and compliance of a certain bar \( f_2 \). The location of the elites are decided from “\( \min f_2(x) \ \text{s.t.} \ f_1(x) \leq 0.1 \text{or} \ f_1(x) \leq 0.2 \text{ or} \ f_1(x) \leq 0.7 \)”.

\[
\begin{align*}
\text{Min } f_1(x) &= 0.04(x_1 \sqrt[3]{x_1^3 + 16} + x_2 \sqrt[3]{x_2^3 + 1}) \\
\text{Min } f_2(x) &= 0.2 \sqrt[3]{x_1^3 + 16} \\
\text{subject to } g_1(x) &= 1 - f_1(x) \geq 0 \\
g_2(x) &= 1 - f_2(x) \geq 0 \\
g_3(x) &= \frac{0.8(x_1^2 + 1)}{x_2 x_3} \\
0 &\leq x_1, x_2 \leq 20, \ 1 \leq x_3 \leq 3
\end{align*}
\]

Figure 6: 2-bar truss problem

3. **Speed reducer problem** is also an engineering design problem formulated below (Fig.7). The aim of this problem is to minimize \( f_1 \), the volume of total material and \( f_2 \), compliance of axis 1. The location of the elites are decided from “\( \min f_1(x) \ \text{s.t.} \ f_2(x) \leq 1300 \text{ or} \ f_2(x) \leq 750 \text{ or} \ f_2(x) \leq 700 \)”.

\[
\begin{align*}
\text{Min } f_1(x) &= 0.7854x_1 x_2 \left( \frac{10 x_3^2}{3} + 14.933 x_3 - 43.0934 \right) - 1.508 x_1 (x_3^2 + x_7^2) \\
&\quad + 7.447(x_5^2 + x_6^2) + 0.7854(x_4 x_6^2 + x_3 x_7^2) \\
\text{Min } f_2(x) &= \sqrt{\left(\frac{745 x_1}{x_2 x_3}\right)^2 + 1.69 \times 10^7} \\
\text{subject to } g_1(x) &= x_1 x_2 x_3 - 27 \geq 0 \\
g_2(x) &= x_1 x_2 x_3^2 - 379.5 \geq 0 \\
g_3(x) &= x_2 x_3 x_6^4 - 1.93 x_4^3 \geq 0 \\
g_4(x) &= x_2 x_3 x_7^4 - 1.93 x_5^3 \geq 0 \\
g_5(x) &= 40 - x_2 x_3 x_4^2 \geq 0 \\
g_6(x) &= 12 x_2 - x_1 \geq 0 \\
g_7(x) &= x_1 - 5 x_2 \geq 0 \\
g_8(x) &= x_4 - 1.5 x_6 - 1.9 \geq 0 \\
g_9(x) &= x_3 - 1.5 x_7 - 1.9 \geq 0 \\
g_{10}(x) &= 1300 - f_2(x) \geq 0 \\
g_{11}(x) &= 110 x_7^2 - \sqrt{\left(\frac{745 x_5}{x_2 x_3}\right)^2 + 1.575 \times 10^8} \geq 0 \\
2.6 &\leq x_1 \leq 3.6, \quad 0.7 \leq x_2 \leq 0.8, \quad 17 \leq x_3 \leq 28 \\
7.3 \leq x_4, x_5 \leq 8.3, \quad 2.9 \leq x_6 \leq 3.9, \quad 5.0 \leq x_7 \leq 5.5
\end{align*}
\]

Figure 7: Speed reducer problem
4. **DTLZ8 problem** is another popular test problem and formulated as follows. We place the three elites at \((f_1, f_2, f_3) = \{(0.0, 0.0, 1.0), (0.72, 0.28, 0.0), (1.2, 0.8, 0.0)\}.

\[
\begin{align*}
\min & \quad f_1(x) = \frac{1}{10} \sum_{i=1}^{10} x_i, \quad f_2(x) = \frac{1}{20} \sum_{i=1}^{20} x_i, \quad f_3(x) = \frac{1}{10} \sum_{i=1}^{10} x_i^3 \\
\text{subject to} & \\
g_1(x) &= f_3(x) + 4f_1(x) - 1 \geq 0 \\
g_2(x) &= f_3(x) + 4f_2(x) - 1 \geq 0 \\
g_3(x) &= 2f_3(x) + f_1(x) + f_2(x) - 1 \geq 0 \\
0 &\leq x_i \leq 1, (i = 1, \ldots, 30)
\end{align*}
\]

Since NSGA II is shown good at providing a well-spread front from the foregoing experiments, we applied it while expecting such performance will still remain even for the constrained problems. All over the problems, we set the parameters such as generation number: 100, elite number: 3, cross-over rate: 0.75 and mutation rate: 0.03. Regarding the population size, it is set at 100 for DTLZ8 and at 50 for the other methods. Moreover, we derived the elites by solving the parametric problems of the \(\varepsilon\)-constraint method. For this, we are ready for using certain powerful software of NLP. Actually, we successfully applied commercial software named LINGO (v.13.0) that can solve large and complicated constrained problems efficiently.

![Figure 8: Comparison of performance between the versions in TNK problem](image1)

![Figure 9: Comparison of performance between the versions in 2-bar truss](image2)

![Figure 10: Comparison of performance between the versions in speed reducer problem](image3)

![Figure 11: Comparison of performance between the versions in DTLZ8 problem](image4)
We compare the results between the original and E-versions in Figs. 8-11. In every case of those constrained problems, we know E-version is apparently superior to the original one, i.e., more precise while dispersing widely and uniformly as well. From all of these results, we can claim the significance of our hybrid method that enables us to gain the merits and remove the demerits of the respective approach shown in Table 1.

5 CONCLUSIONS

As a promising decision aid for problem-solving in global and competitive manufacturing, this study concerned with multi-objective analysis (MOA) and evaluated an elite-induced evolutionary algorithm for MOA (EIEA/MOA) from comprehensive point of view. For this purpose, we have deployed the idea over several MOEA methods such as PSA, MOPSO, MODE, NSGA II and SPEA II and compared the manifold properties among these variants in terms of a few metrics defined originally. Finally, to discuss on the performances in practice, numerical experiments have been carried out based on a simple test problem and the constrained problems. Through elaborate discussions from the numerical results, we ascertained the effectiveness of the proposed EIEA/MOA and revealed some properties regarding these variants. Significant attentions on future studies should be turned to providing more efficient procedures for strong induction and the post-optimal analysis of MOP relying on EIEA/MOA. Examining the performance for many-objective problems [18,19] is also interesting as a hot topic in this area.

List of Abbreviation

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR</td>
<td>Average rank</td>
</tr>
<tr>
<td>EIEA</td>
<td>Elite Induced Evolutionary Algorithm</td>
</tr>
<tr>
<td>MOA</td>
<td>Multi-objective Analysis</td>
</tr>
<tr>
<td>MOEA</td>
<td>Multi-Objective Evolutionary Algorithm</td>
</tr>
<tr>
<td>MOPSO</td>
<td>Multi-objective Particle Swarm Optimization</td>
</tr>
<tr>
<td>NSGA II</td>
<td>Non-dominated Sorting Genetic Algorithm</td>
</tr>
<tr>
<td>POSS</td>
<td>Pareto Optimal Solution Set</td>
</tr>
<tr>
<td>SP</td>
<td>Spacing;</td>
</tr>
<tr>
<td>DM</td>
<td>Decision Maker</td>
</tr>
<tr>
<td>IR</td>
<td>Induction rate</td>
</tr>
<tr>
<td>MODE</td>
<td>Multi-Objective Differential Evolution</td>
</tr>
<tr>
<td>MOP</td>
<td>Multi-Objective Optimization</td>
</tr>
<tr>
<td>NLP</td>
<td>Nonlinear programming</td>
</tr>
<tr>
<td>PSA</td>
<td>Pareto Simulated Annealing</td>
</tr>
<tr>
<td>SA</td>
<td>Simulated Annealing</td>
</tr>
<tr>
<td>SPEA II</td>
<td>Strength Pareto Evolutionary Algorithm II</td>
</tr>
</tbody>
</table>

REFERENCES


THERMAL-ECONOMIC DESIGN OF A MICRO-GAS TURBINE CHP SYSTEM USING A MULTI-OBJECTIVE OPTIMIZATION APPROACH

Ana C. Ferreira¹, A. Ismael F. Vaz¹, Senhorinha F. Teixeira¹ and José C. Teixeira²

¹Department of Production and Systems, School of Engineering, University of Minho
Azurém, 4800-058 Guimarães, Portugal
{acferreira; aivaz; st}@dps.uminho.pt

²Department of Mechanical Engineering, School of Engineering, University of Minho
Azurém, 4800-058 Guimarães, Portugal
jt@dem.uminho.pt

Key words: Micro-gas turbines, Multi-objective Optimization, Cogeneration Systems.

Summary. Micro gas-turbines are addressed as an option for decentralized energy generation. The reasons behind the use of micro-turbines include the technologic maturity stage, multi-fuel capability, low emission levels, heat recovery potential, and need of minimal maintenance. Therefore, optimization of such systems is of the greatest importance in the area of power generation. Combined heat and power (CHP) systems are considered the best option to produce both heat and power simultaneously due to increased efficiency and reduced energy cost. In this paper, it is presented a multi-objective optimization of a micro gas turbine operating in cogeneration mode. The system aims to provide electric and thermal power to fulfil the energy demands for a reference case, which corresponds to block of apartments located in Portugal. This CHP system is composed of air compressor (C), combustion chamber (CC), an internal air pre-heater (IPH), a gas turbine (GT), and a water heat exchanger (WHE). Each component of the micro turbine cycle is modelled using the energy balances of the first law of thermodynamics. It is also proposed an economic model that defines the purchase cost of each system component. The cost equations include thermodynamic variables that directly affect the component cost and performance. The mathematical model yields two non-linear objective functions: the minimization of the total investment cost of the system and the maximization of the overall efficiency of the system. The optimization problem was solved by NSGA II algorithm. The multi-objective optimization results were expressed by Pareto curves. The optimal solutions disclose a set of optimal solutions for which the CHP system has overall efficiencies between 77.8% and 85.0% whereas the annualized investment costs varies between 12 585 €/year and 26 654 €/year.

1 INTRODUCTION

Recently, decentralized power generation resulted in remarkable outputs concerning the decrease in energy waste and the increase in its reliability. Local small and micro-scale power plants will further contribute in power generation, and it is well believed that in the not too distant future, they will replace large and centralized power plants to fulfil the energy demands [1–3].
Micro gas-turbines (in a range of 25-250 kWel) are addressed as one of the most serious options for decentralized energy generation and they offer a number of potential advantages compared to other technologies for small-scale power generation. The use of micro gas-turbines, as one of the main technologies applied in energy generation relies on higher efficiency, lower commissioning time, compact size and low-weight per unit of power, leading to reduced civil engineering costs, higher runtime, and lower maintenance (due fewer movable parts), multi-fuel capabilities, as well as, opportunities for lower emissions, mainly in the Combined Heat and Power systems (CHP) operation mode [4,5]. In addition, gas turbines enjoy certain merits relative to diesel engines in the context of mini- and micro-power generation. They have high-grade waste heat, low vibration level and short delivery time. Nowadays, the main barriers to the implementation of CHP systems are the great investment costs and high payback period, the costs of grid connection, the access to the electricity network (that sometimes is not always possible, concerning the location or even the law barriers in some countries). Micro gas-turbines could be used for power generation in the industrial, commercial and residential sectors but one of the largest uses for micro-turbines is in CHP, allowing to produce both heat and power simultaneously, and as a consequence, increasing overall efficiency and reducing the energy costs generation [6].

Technically, the exhaust heat from a gas turbine is of high quality i.e. high temperature and it can be used to produce heat for space heating or other thermal uses. For micro gas-turbines, in order to get the high efficiencies the exhaust gas is passed through a recuperator (also called internal pre-heater) to increase the electrical efficiency and therefore the final heat grade available is of lower quality (temperature). The turbine inlet temperatures are low (1075–1275 K) to use relatively inexpensive materials for the turbine, keeping the costs at a reasonable level and maintaining low-pressure ratios (3.0 to 4.0). The low inlet temperature affects the efficiency, which can be counteracted by using recuperation [7].

Additional value in cogeneration mode is gained when the thermal energy from the exhaust gases is recovered to supply local heat. Commonly, an integrated heat exchanger is used to extract heat from the exhaust gas before its release to the atmosphere. Depending on the application, hot water in the 70–90 ºC temperature range may be produced. As a result, micro gas-turbines typically apply a recuperated cycle and achieve an overall efficiency exceeding 75-85% (based on Low Heating Values, LHV) [8].

The power produced by the turbine and consumed by the compressor is proportional to the absolute temperature of the gas passing through them. Thus, it is advantageous to operate the expansion turbine at the highest practical temperature consistent with economic materials. With the technological advance and the use of more resistant materials, it is possible to get higher turbine inlet temperatures. Thus, the trend in gas turbine advancement has been towards a combination of higher temperatures and pressures, which increase the production costs and the technology investment costs. The specific investment cost potentially increase with the decrease of the cogeneration system size as can be seen in Figure 1. In Figure 1, it is presented the specific investment cost and respective curve fitting, as a function of the power plant size for internal combustion engines and gas turbines [9].

For this reason, micro gas-turbines and other distributed energy technologies to be competitive in power markets, the price of electricity production will need to be more attractive than today. Thus, optimization of such systems combining energy and cost aspects is of the greatest importance in the area of power generation.
Energy utilization is essentially governed by thermodynamic principles. Thermal-economic analyses combines the energy balances with the economic principles and incorporates the total cost of an energy system. These costs can conduct designers to understand the relationship between the gains in the energy system performance and its increase effect in the production cost. Numerous researchers, e.g. [10–16] have conducted thermal-economic analysis and optimization for thermal systems in two different streams: (i) studies focused on the cost flow analysis if the objective is the determination of cost generation and/or cost losses and (ii) studies focused on the optimization of the design, i.e., the selection of the best system operational conditions if the objective is the optimization of power system. It becomes essential to use the optimization procedure considering thermodynamics laws as well as exergo-economics. In this regard, the main objectives involved in the design optimization process are [17]: thermodynamic (e.g. maximum efficiency, minimum fuel consumption, minimum irreversibility, and so on), economic (e.g. minimum cost per unit of time, maximum profit per unit of production), and environmental (e.g. limited emissions, minimum environmental impact). The purpose of the present work is the multi-objective optimization of a micro gas turbine operating in cogeneration mode considering two non-linear objective functions: the minimization of the total investment cost of the system and the maximization of the overall efficiency of the system. The optimization problem was solved by NSGA II algorithm.

2 SYSTEM DESCRIPTION
The problem presented in this study aims to optimize a small-scale cogeneration system by producing electrical power, and simultaneously be able to fulfil both the heating and the domestic hot water needs, for a building of residential apartments. For this scale of application, the system must operate, approximately, 4000 h which corresponds to a heat output of 125 kW. The building consists of a 52 individual dwellings with an individual floor area of 150 m² (or 7800 m² in total). The annual thermal power duration curve of the building was calculated according to the Portuguese regulation for the thermal behaviour of buildings (RCCTE, Decree Law 78/2006) [25], by summing the hourly heating load and the hourly hot water needs. The domestic hot water needs calculations were performed considering an occupation of 4 people per dwelling with a daily domestic hot water consumption of 40 L per person, at a temperature of 333 K. The building hourly heating loads were calculated considering a class B minus building and the local climate (i.e. north of Portugal).
This study is based on a small-scale cogeneration system designed under the Joule-Brayton cycle. Figure 2 illustrates the layout of a micro turbine based CHP system. The turbo machinery and the electric generator are connected to a common shaft rotating at high speed. An inverter decouples the high frequency of the produced current from that of the grid, thus enabling variable speed operation. For so small applications with a low-pressure compressor, gas turbines require an Internal air Pre-Heater (IPH) or regenerator, to provide a satisfactory electrical efficiency. The thermal energy of the Exhaust Gases (EG) is recovered as useful heat for heating water. Atmospheric air is compressed (C) and fed to the IPH before entering the Combustion Chamber (CC) where it is mixed with Natural Gas (NG).

The high temperature combustion gases expand in the Turbine (T). The EG, leaving the turbine, are firstly used in the IPH to pre-heat the incoming compressed air and subsequently for the production of hot water in the external heat recovery system, before exiting to the atmosphere. The latter is a Water Heat Exchanger (WHE), where a fixed flow rate of water is heated from 313 K to 353 K.

2.1 Physical Model

For the physical model equations, the operating fluid properties were calculated through standard thermodynamic relationships. All the components were considered adiabatic and a degree of irreversibility was assumed for both the compressor and the turbine. The air and gases were treated as perfect gases with constant specific heats and the compressor inlet conditions were assumed as $T_1 = 293$ K and the atmospheric pressure $P_0 = 1.013$ bar.

A reasonable pressure drop is assumed for the flows through the internal pre-heater, combustion chamber and external heat exchanger: $P_5 = 0.95P_3$ and $P_6 = 0.97P_5$ for the internal pre-heater; $P_4 = 0.95P_3$ for the combustion chamber; and $P_7 = 0.95P_6$ for the water heat exchanger, where the $P_i$ and $P_{i+1}$ represent the inlet and the outlet pressure after the pressure drop, respectively. The main relationships associated to the five plant components are listed below by equations (1)-(9). The air compressor inlet temperature, $T_2$, can be calculated by (1),

$$T_2 = T_1 \left[ 1 + \left( \frac{P_{i+1}}{P_i} \right)^{\frac{\gamma - 1}{\gamma}} \right]$$  

(1)
where $\eta_C$ is the compressor isentropic efficiency, $\gamma$ is the polytropic adiabatic coefficient and $r_c$ compressor pressure ratio and thus, $P_2 = r_c \cdot P_1$. The mechanical power absorbed by the compressor is given by, $\dot{W}_c = \dot{m}_a \cdot c_{pa} \cdot (T_2 - T_1)$ and the air mass flow rate, $\dot{m}_a$, by (2),

$$\dot{m}_a = \frac{W}{(1+f) \cdot c_{pg} \cdot (T_4 - T_3) - c_{pa} \cdot (T_2 - T_1)}$$

where $f$ represents the fuel to air mass ratio (defined below) and the $C_{pa}$ and $C_{pg}$ are the specific heats at constant pressure of air and flue gases, respectively, and $\dot{W}$ is the net mechanical power.

The fuel properties are those of the standard natural gas and a combustion efficiency of 98% is considered. The fuel mass flow rate, $\dot{m}_{fuel}$, can be estimated by the relationship, $\dot{m}_{fuel} = \dot{m}_a \cdot f$ with the fuel to air mass ratio given by (3),

$$f = \frac{c_{pg} \cdot (T_a - T_4) - c_{pa} \cdot (T_3 - T_4)}{LHV \cdot 0.98 \cdot c_{pg} \cdot (T_4 - T_3)}$$

The gas mass flow rate, $\dot{m}_g$, resultant from the combustion of air and fuel is simply the sum of the air and fuel mass flow rates, $\dot{m}_g = \dot{m}_a + \dot{m}_{fuel} = \dot{m}_a \cdot (1+f)$.

The turbine pressure ratio, $r_t$, is smaller than $r_c$ due to the successive pressure losses occurring in the combustion chamber and the heat-exchangers, as (4),

$$r_t = \frac{P_3 \cdot P_4}{P_2 \cdot P_5} = r_c \left( \frac{P_3 \cdot P_4}{P_2 \cdot P_5} \right)$$

The exhaust gases, leaving the combustion chamber at the controlled maximum temperature $T_4$, are then expanded in the turbine and its exit temperature ($T_5$) is estimated by the relationship, $T_5 = T_4 \left( 1 - \eta_T \cdot (1-r_t^{0.5}) \right)$ where $\eta_T$ is the turbine isentropic efficiency. The mechanical power delivered by the turbine is given by, $\dot{W}_T = \dot{m}_g \cdot c_{pg} \cdot (T_4 - T_5)$.

As previously mentioned, the system includes an internal pre-heater to increase the micro gas-turbine electrical efficiency. This plant component is basically an air-gas heat-exchanger used to pre-heat the air before entering the combustion chamber. For a given heat-exchanger effectiveness $\epsilon_{IPH}$, the temperature ($T_3$) can be calculated by $T_3 = T_2 \left( 1 - r_t^{0.5} \right) + \Delta T_{IPH \cdot \epsilon_{IPH}}$. In the heat exchange between the air ($T_2$) and the expanded gases ($T_3$), it is assumed that the thermal energy of the hot gases is fully transferred to the air flow. Based on the inlet and outlet temperatures of the two streams, the heat-exchanger logarithmic mean temperature difference, $\Delta T_{IPH}$, is defined as (5),

$$\Delta T_{IPH} = \frac{(T_5 - T_3) - (T_6 - T_2)}{\ln \left( \frac{T_5 - T_3}{T_6 - T_2} \right)}$$

The required heat transfer area, $A_{IPH}$ can then be calculated by (6),
\[ A_{IPH} = \frac{m_g c_{pg} (T_5 - T_6)}{U_{IPH} \Delta T_{\ln,IPH}}, \]

where \( U_{IPH} \) is the overall heat transfer coefficient in the IPH. The hot gases are used to heat water through an external WHE. A required thermal power \( \dot{Q} \) to heat the water stream from \((T_5)\) to \((T_9)\) is required and it can be defined as, \( \dot{Q} = m_w c_{pw} (T_9 - T_8) \), where the water mass flow \( m_w \) is directly calculated and the exhaust-gasses temperature \((T_7)\) can be obtained from the adiabatic heat-balance of the heat-exchanger, as \( (7) \),

\[ T_7 = T_6 - \left[ \frac{m_w c_{pw} (T_9 - T_8)}{m_g c_{pg}} \right], \]

where \( c_{pw} \) is the specific heat of the water.

Likewise the IPH, the logarithmic mean temperature, \( \Delta T_{\ln,WHE} \), and the required heat transfer area, \( A_{WHE} \) are calculated as in \( (8) \) and \( (9) \) respectively,

\[ \Delta T_{\ln,WHE} = \frac{(T_6 - T_o) - (T_7 - T_o)}{\ln \left( \frac{T_6 - T_o}{T_7 - T_o} \right)}, \]

\[ A_{WHE} = \frac{m_g c_{pg} (T_6 - T_7)}{U_{WHE} \Delta T_{\ln,WHE}}, \]

where \( U_{WHE} \) is the overall heat transfer coefficient in the WHE.

Finally, the net mechanical power delivered by the mechanical shaft of the turbine-compressor \( \dot{W} \), is the difference between the mechanical powers produced by the turbine and absorbed by the compressor, \( \dot{W} = \dot{W}_t - \dot{W}_c \) and the electrical power \( \dot{W}_el \) includes the efficiency of the electrical generator, herein assumed as constant and equal to 93\%, i.e., \( \dot{W}_el = 0.93 \cdot \dot{W} \).

### 2.2 Economic Model

The mathematical expressions that define the cost for each component of the CHP system, were based on previous studies [8,18], where the exponents and coefficients of each cost equation were adjusted for small-scale units, also taking into account real market data, by using a micro gas-turbine available in the market as a reference. As a result, the cost estimation can be performed in order to evaluate the overall cost of the system for a specific range of combined power production. The purchase cost equation for each component, \( C_{Cc}, C_{Cc}; C_{IT}, C_{IT}; C_{IPH}, C_{WHE} \), is presented in Table 1.

Each purchase cost equation was defined considering physical variables that integrate the thermodynamic model. These variables can be divided in size and quality variables. The terms \( C_{11} \) to \( C_{53} \) represent the respective cost coefficients taking into account the reference prices and the relative weight of each component in the total investment cost of the CHP system. Their values and validation were presented and discussed elsewhere [8].
<table>
<thead>
<tr>
<th>Component</th>
<th>Purchase Cost Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Air Compressor</td>
<td>$C_c = C_{11} \left( \frac{\dot{m}<em>a}{\dot{m}</em>{a,ref}} \right)^{0.8} r_c \ln(r_c) \frac{0.92 - \eta_c}{0.92 - \eta_c}$</td>
</tr>
<tr>
<td>Combustion Chamber</td>
<td>$C_{cc} = C_{21} \left( \frac{\dot{m}<em>a}{\dot{m}</em>{a,ref}} \right)^{0.8} (1 + e^{C_a(T_c - C_{a1})}) \frac{0.995 - P_3}{P_3}$</td>
</tr>
<tr>
<td>Turbine</td>
<td>$C_t = C_{31} \left( \frac{\dot{m}<em>g}{\dot{m}</em>{g,ref}} \right)^{0.8} (1 + e^{C_a(T_r - C_{a1})}) \ln(r_r) \frac{0.92 - \eta_r}{0.92 - \eta_r}$</td>
</tr>
<tr>
<td>IPH</td>
<td>$C_{IPH} = C_{41} A_{IPH}^{0.4} (1 + e^{C_a(T_r - C_{a1})})$</td>
</tr>
<tr>
<td>WHE</td>
<td>$C_{WHE} = C_{51} A_{WHE}^{0.4}$</td>
</tr>
</tbody>
</table>

### 3 MULTI-OBJECTIVE OPTIMIZATION VIA NSGA II ALGORITHM

In various domains of engineering, several real-world optimization problems are extremely difficult and complex in terms of number of variables, nature of the objective function, existence of many local optimal, or imply a continuous or discrete search space. Practical engineering application, as the one presented herein, usually consider more than one conflicting objective function to be optimized, resulting in multi-objective optimization problems. Usually, a unique solution that simultaneously minimizes all the objective function does not exist and, therefore, different solutions will produce trade-offs between different objectives values. The set of solutions create a multi-dimensional function space, typically represented by Pareto curves [19].

#### 3.1 Multi-objective algorithm

An algorithm based on non-dominated sorting was proposed by Srinivas and Deb and called non-dominated sorting genetic algorithm (NSGA). This was later modified by Deb et al. [20] who eliminated higher computational complexity and the need for specifying additional parameters. Thus, NSGA II turned a popular non-domination based genetic algorithm for multi-objective optimization, which perfectly suits our problem characteristics, since NSGA II can also address non-linear constraints. The optimization problem was solved considering 200 generations with a population size of 100 for the NSGA II algorithm.
3.2 Objective functions, decision variables and non-linear constraints

Objective Functions

In the present study, two key objective functions were defined for optimization:

- Minimization of the total investment cost of the CHP system, $\text{Min } C_{\text{inv}}$, which is annualised as if it were being paid off a loan at a particular interest of discount rate over the estimated CHP system lifetime. So, $C_{\text{inv}}$ represent the sum of purchase cost equations presented in Table 1, multiplied by the capital recovery factor (CRF), as defined by equation (10),

$$
C_{\text{inv}} = CRF \cdot \left( C_C + C_{CC} + C_T + C_{IPH} + C_{WIE} \right)
$$

(10)

where CRF is used to determine the equal amounts of $A$ cash transactions for an investment considering the effective rate of return, $i_e = 7\%$, over $n$ years of the lifetime (10 years), as in equation (11).

$$
CRF(P \rightarrow A, i_e, n) = \frac{i_e(1+i_e)^n}{(1+i_e)^n - 1}
$$

(11)

- Maximization of the overall efficiency of the cogeneration system, $\text{Max } \eta_{\text{overall}}$, defined in equation (12).

$$
\eta_{\text{overall}} = \frac{W_{\text{el}} + Q}{m_{\text{fuel}} \times \text{LHV}}
$$

(12)

The overall efficiency of a cogeneration system is the best way to measure and compare the cogeneration performance with the conventional power production. The total efficiency of a system is measured as the fraction of the input fuel that can be converted in usefully power and heat. Most manufacturers of this kind of power plants relate efficiency to the amount of fuel that is consumed and the respective lower heating value of the input fuel (LHV).

Decision Variables

The decision variables (design parameters) in this study are compressor pressure ratio ($\tau_c$), compressor isentropic efficiency ($\eta_c$), turbine isentropic efficiency ($\eta_t$), combustion chamber inlet temperature ($T_3$), turbine inlet temperature ($T_4$) and the amount of electrical power produced ($W$). Variables are further constrained to be within reasonable ranges, and, so, simple bounds have been defined, as presented in Table 2. The table also presents the justification regarding decision variable choice.

Non-linear Constraints

Regarding the definition of non-linear constraints aims to restrict the implicit variables of the problem according to their physical significance in the system operation. The non-negativity condition for all the temperatures over the thermodynamic cycle were assumed as constraints in the model.
Table 2: Definition of simple bounds for the decision variables

<table>
<thead>
<tr>
<th>Decision Variable</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>$3 \leq r_c \leq 10$</td>
<td>Commercial availability</td>
</tr>
<tr>
<td>$0.7 \leq \eta_c \leq 0.9$</td>
<td>Commercial availability</td>
</tr>
<tr>
<td>$0.7 \leq \eta_t \leq 0.9$</td>
<td>Commercial availability</td>
</tr>
<tr>
<td>$500 \leq T_3 \leq 1000$</td>
<td>Thermodynamic relationships</td>
</tr>
<tr>
<td>$1000 \leq T_6 \leq 1400$</td>
<td>Material temperature limit</td>
</tr>
<tr>
<td>$90 \leq \dot{W} \leq 125$</td>
<td>Keep the heat-to-power ration within the specific values for micro-gas turbines, $\lambda \leq 1.25$</td>
</tr>
</tbody>
</table>

The high-pressure air is pre-heated before entering in the CC and so it is required that the temperatures $T_2$ should be lower than the temperature $T_3$, ($T_2 \leq T_3$). On the other hand, $T_3$ must be lower than the exhaust gases temperature ($T_5$) at the turbine exit, in order to guarantee an effective heat transfer in the IPH, ($T_3 \leq T_5$). These physical limitations should also satisfy the temperature bounds: $50 \leq T_3 - T_2 \leq 600$ and $10 \leq T_5 - T_3 \leq 700$. Similarly, the inlet temperature of the air must be lower than the air temperature after suffering the pre-heating in the IPH ($T_3$) and, therefore, the latter must be lower than the temperature of the combustion gases at the turbine inlet ($T_4$), ($T_3 \leq T_4$) which is the highest temperature reached in the system. Also, the difference between the inlet and outlet temperatures in each of heat exchangers flows should be bounded to ensure the effectiveness in the process of heat transfer between the fluids: $10 \leq T_6 - T_2 \leq 800$ and $200 \leq T_6 - T_5 \leq 800$. The temperature $T_7$ cannot be lower than 363K in order to prevent the condensation problems in the heat recovery system, $363 \leq T_7 \leq 1000$. Primary energy savings (PES) [21] was also included in the model as an inequality constraint in order to guarantee that the system may be classified as high-efficient CHP power plant, which corresponds to a value above 10% for small-scale applications ($PES \geq 0.1$).

4 RESULTS AND DISCUSSION

The multi-objective optimization results are expressed as a two-dimensional Pareto curve (front) in Figure 3. In multi-objective optimization, each of the points in Pareto curve represents a point where one objective function value may not be improved without worsening others. Therefore, decision-makers select one of them based on the best interest conditions. Decision-making process is usually performed with the help of an assumed (ideal) point (point B in the Figure 3) formed by the individual objective functions optimal solution. Observing Figure 3 and noting that the ideal point is not part of the Pareto front we conclude that the objective functions considered are in conflict. Please note that we are minimizing the equipment purchase investment cost while the overall efficiency is being maximized.

As shown in Figure 3, the optimal solutions disclose a set of optimal solutions for which the CHP system has overall efficiencies between 77.8% (point A in Figure) and 85.0% (point...
C in Figure 3) and annualized investment costs between 12,585 €/year (point A in Figure 3) and 26,654 €/year (point C in Figure 3). From the analysis of the Pareto curve, it is shown that gains in overall efficiency above 83% imply a great increase in the total CHP purchase cost. The nearest point on Pareto curve to “ideal” point may be regarded as the “desirable” solution (region D in Figure 3). Nevertheless, the selected point depends on the balance of the objective functions changes.

![Figure 3: Pareto optimal frontiers in objectives’ space.](image)

In the solution variables space, Figure 4 represents the dispersion of the six decision variables for the optimal solutions set grouped within two 3-dimensional projections. From the analysis, the set of optimal solutions give priority to a system with a power output close to 90-91 kW (decision variable lower limit) and constant isentropic efficiencies ($89\% \leq \eta_C \leq 90\%$ and $86\% \leq \eta_T \leq 88\%$).

![Figure 4: Distribution of the six decision variables for the objectives’ space: a) representation of $r_C$, $\eta_C$ and $T_3$; b) representation of $T_4$, $\eta_C$, and $W$.](image)

Nevertheless, the most critical decision variables seem to be the compressor pressure ratio, turbine inlet temperature, and the combustion chamber inlet temperature. This particular analysis can give an insight regarding the decision variables selection, and so, it is possible that in further studies the compressor and turbine isentropic efficiencies to be assumed as constant values.
Is this study it was also performed a sensitivity analysis regarding changes on different parameters: an aspect with an economic impact, i.e., the CHP system lifetime expectancy; an aspect with thermodynamic implication, i.e., the air inlet temperature admitted to the CHP system. This latter has an increased relevance because of temperature changes throughout the seasons of the year (Figure 5).

Figure 5 a) shows sensitivity of Pareto curve for different values of the operational systems lifetime (10, 15, and 20 years). This analysis shows that Pareto curve is relocated in terms of investment cost with the increase of system lifetime, but the overall efficiency do not present any differences. Figure 5 b) shows sensitivity of Pareto curve for different values of air inlet temperature (288 K, 293 K, and 298 K). This parameter seems to affect both objective functions, since for higher values of inlet air temperature there is a decrease in the CHP purchase investment cost for higher values of overall efficiency.

4 BRIEF CONCLUSIONS

In this paper, thermo-economic modelling and multi-objective optimization of a micro gas-turbine have been done. A multi-objective optimization was performed by defining as objective functions the maximization of overall efficiency and the minimization of total purchase cost of system. Pareto curves were obtained and analysed. The obtained Pareto curve disclosed several design possibilities for which the overall efficiencies vary between 77.8% and 85.0% for an annualized investment costs fluctuating between 12 585 €/year and 26 654 €/year.

This study led to the conclusion that the compressor pressure ratio, turbine inlet temperature, and the combustion chamber inlet temperature are the most relevant decision variables. The NSGA II algorithm seems to be adequate to solve this thermal-economic problem. However, further sensitivity analyses should be performed in order to test other algorithm options or objective functions allowing significant insights about the problem.

ACKNOWLEDGES

The authors are grateful for the support provided by FCT national funds under the Strategic Project PEst2015-2020, with the reference: UID/CEC/00319/2013.
REFERENCES

LOGISTICS OF TRANSITION FROM WORKING STRESS TO LIMIT STATES DESIGN IN FOUNDATIONS ENGINEERING

KONSTANTIN G. ASHKINADZE ¹

¹ Jacobs Engineering Ltd.
#1800, 10065 Jasper Avenue
Edmonton, Alberta, Canada
konstantin.ashkinadze@jacobs.com
www.k-the-eng.com

Key words: Limit State Design, Working Stress Design, Global Safety Factor, Load Factor, Resistance Factor, Pile Foundation.

Abstract. The paper considers transition from Working Stress Design (WSD) using a single Global Safety Factor (GSF) to Limit States Design (LSD) using differentiated load and resistance factors in one specific aspect of foundations engineering – design of pile foundations for vertical compressive loads. Conversion to LSD format in Canadian design codes in 2005 did not assure consistency of designs with the earlier WSD paradigm; the equivalent GSF has increased from 2.5 to 3.6. Due to it, the piles for the same loads became longer or larger than before, and consequently more expensive. This undermines credibility of the LSD paradigm in the eyes of the practitioner since the existing WSD designs have performed adequately.

Using this case as an example, the paper examines a more general question of probabilistic calibration of load and resistance factors in LSD and similar methods. It is demonstrated that uniform level of reliability is not achieved in these methods. The imposed reliability of resistance is generally much higher than the reliability of load calculation, which limits usefulness of these methods for direct probability-of-failure-based design. The results of probabilistic calculations critically depend on the assumed coefficient of variation of both the load and the resistance, which can only be back calculated from data for existing designs.

It is argued that uncertainties inherent in geotechnical problems do not allow any more refined quantification of design reliability than that afforded by the single number of GSF. In the process of transition from WSD to LSD, it is suggested to maintain the compatibility condition GSF = α / φ (α > 1 – load factor, φ < 1 – resistance factor) in all cases, using probabilistic considerations only for information and control. Code modification should only occur if the results of new tests deviate significantly from the existing data, or the amount of the new test data becomes comparable with the tremendous number of foundations designed under WSD and performing adequately.

1 INTRODUCTION

The traditional Working Stress (WSD) – a.k.a. Allowable Stress (ASD) – design approach in structural, foundations and geotechnical engineering is being phased out worldwide and replaced with the more advanced Limit States Design (LSD) paradigm. The primary driver
behind this transition is the urge to replace the intuitive process of assignment of safety margins based on experience with the more orderly probabilistic treatment based on the actual risk of failure.

The WSD lumped “everything that may go wrong” about a particular mode of failure, or sometimes even across several modes of failure, into a single Global Safety Factor (GSF). Even in the more clearly defined case of a single mode of failure, the GSF combined the uncertainty of load determination, resistance determination, material properties variations, dimensional deviations etc. etc. Although the GSF implicitly took into consideration the degree of scatter in the analytical data as evidenced and corroborated by experience, it was not expressly linked to any probabilistic criteria.

The LSD came with a promise to create a rational basis to the assessment of the structural reliability, which was necessitated by insurance and liability considerations most of all, because really there was nothing wrong with the designs produced using the traditional WSD approach. Another LSD promise was attainment of more economical designs because in the ultimate limit state, the structure may exhibit additional reserves of strength. The WSD controlled achievement of the “allowable stress” value in one point only; the LSD could allow it to develop in a contiguous region of the structure (e.g. “yield line” or “plastic hinge”), as dictated by the considered failure mechanism. This was not an inherent advantage, though, because the WSD could also be formulated to consider the “ultimate” condition, which would then be relaxed by the GSF. The true difference between these two methods is the assignment of the safety factors. The design formula of the WSD, in case of only one load, is

\[ L \leq \frac{R}{GSF} \]  

where \( L \) = the load effect and \( R \) = the resistance. The design formula for the LSD is

\[ \alpha L \leq \phi R \]  

where \( \alpha \) = the load factor (> 1) and \( \phi \) = the resistance factor (< 1). Comparing Equations (1) and (2), immediately obtain the condition that must be satisfied for the WSD and LSD approaches to produce identical designs:

\[ GSF = \frac{\alpha}{\phi} \]  

The LSD paradigm provided the framework for assignment of the factors \( \alpha \) and \( \phi \) from probabilistic considerations, such as non-exceedance of a certain multiple of the “specified” load or the resistance not falling below a certain fraction of the “ultimate” value. (The assignment of the “specified” value of \( L \) and the “ultimate” value of \( R \) is a separate problem, which arises in both the WSD and LSD paradigms.) However, trying to rationalize about the nature of “all perils” that have previously been expressed by the single GSF is itself perilous. The uncertainties in engineering designs are ordinarily classified as aleatory and epistemic [1]. Aleatory uncertainties are ones that are caused by the unavoidable randomness of the phenomena considered. Epistemic uncertainties are ones that originate from lack of data or insufficient available knowledge, and can be reduced through accumulation of additional data or knowledge.
In this sense, the WSD treated all uncertainties encountered in the design as aleatory, whereas the LSD treats them as epistemic. It implies that by additional measurements of loads and resistances, it is possible to rigorously calculate the probability of failure and control it by setting the appropriate thresholds $\alpha$ and $\phi$.

In reality, both types of uncertainty are present in any engineering design. The WSD did not allow rationalizing and controlling the probability of failure, and this is the reason why it was abolished. The LSD came with the optimism that this control is possible, at least in principle, despite the ever-present aleatory uncertainties. However, even the optimists had to concede that no amount of experimentation in the foreseeable future will be able to match the quiet process of data accumulation that has gone on for decades, producing millions of successfully completed structures designed under the WSD approach. The only information available about them was that they worked; hence the GSF values used in their design must have been adequate in the face of all kinds of uncertainty. Therefore, the only practical way of calibrating the $\alpha$ and $\phi$ factors was to firmly uphold Equation (3). It made the LSD work in the early stages but it was also an admission of failure, because so far the uprooting of the traditional WSD method has not led to any improvements in design reliability.

2 CALIBRATION OF LSD: WSD VS. PROBABILITY

The calibration of resistances of structures made of manufactured materials (steel, concrete, masonry, engineered lumber) or naturally occurring but gradable and selectable material (solid sawn lumber) was facilitated by quality control and factory testing, which allowed accumulating reliable statistics. It still did little to alleviate the aleatory uncertainty, so Equation (3) still needed to be enforced, but it helped to link the resistance factors with probability and reliability concepts. This gave encouragement to the specialists that finally, the promise of directly considering the probability of failure was being delivered by the LSD.

In the early 2000s, the LSD was adopted as the design paradigm of choice in geotechnical and foundations engineering in Canada [2]. Prior to this, the geotechnical engineering community largely resisted the LSD, stating that “soil is not steel” and that the uncertainties that abound in geotechnical problems did not allow any more refined quantification than the single number of GSF. However, it was becoming apparent that the design procedure for foundations had to be harmonized with that used for the design of the superstructure.

The resistance factors for different types of foundation and geotechnical problems in [2] were established based on the works by Becker [4, 5]. These factors are being used in the current Canadian codes [3, 6], therefore our review will largely be focused on Becker’s derivations. Subsequent works in this field in Canada and the USA [8, 9, 10] elaborated on further linking the “factored” resistance of foundations (deep foundations, in this case) to the statistics of newly available pile tests, thereby driving away from Equation (3) and the legacy of the WSD method.

In shallow foundation design using WSD design paradigm [7], the GSF was routinely taken as 3.0. No distinction was made whether this was caused by ultimate (plunging) shear failure of the base soil, or by excessive settlement. For deep foundations in initial stages of design, the GSF was usually 2.5, reducing to 2.2 at the final stages of design. When pile testing data were available, the GSF could be reduced to 2.0 and even 1.8.

With the advent of LSD, the resistance factors for shallow foundations were calibrated as
0.5; for deep foundations, they were set as 0.4 in compression and 0.3 in tension if the soil properties were obtained by correlation or laboratory tests. If the actual static pile test data were available, the resistance factors would become 0.6 for compression and 0.4 for tension. Note that [7] suggested that the allowable stress in tension in WSD be taken as 50% that for compression, although it is often omitted in actual geotechnical reports. In the LSD paradigm, according to [6], the pile resistance in tension should be taken as 70-80%, and possibly even 50% (for piles with high driving stresses) of their compressive resistance. This is accounted for by setting a lower resistance factor for piles in tension. Applying Equation (3) and assuming \( \alpha = 1.45 \) (typical value for residential and commercial properties under the Canadian code [3] when Live load is 4 times greater than Dead load), find the following “equivalent” GSF values:

- Shallow foundations: GSF = 1.45/0.5 = 2.9 – close to 3.0;
- Deep foundations, compression, no static test: GSF = 1.45/0.4 = 3.6 – not close to 2.5;
- Deep foundations, tension, no static test: GSF = 1.45/0.3 = 4.8 vs. 2.5/0.5 = 5.0 in WSD – close;
- Deep foundations, compression, with static test: GSF = 1.45/0.6 = 2.4 – significantly different from 2.0;
- Deep foundations, tension, with static test: GSF = 1.45/0.4 = 3.6 vs. 2.0/0.5 = 4.0 in WSD – reasonably close.

As appears from this summary, the resistance factors assessed for pile foundations loaded in compression without direct static test data are not calibrated consistently with the historical WSD design. The new LSD paradigm is much more conservative than the old WSD methodology. (The same would be true for piles in axial tension if the 50% reduction of the allowable capacity prescribed by [7] were not applied, as is often the case; for example, the classical text [11] makes no mention of this reduction.) Since compression piles designed by correlation of soil shear strength with Standard Penetration Test (SPT), pocket penetrometer, laboratory compressive tests and other indirect methods account for the majority of piles installed in construction practice, this loss of design consistency has caused major problems for the piling industry in Canada. The piles for the same loads have become longer or larger than before, and therefore more expensive. In revamp projects where existing piles are reused, they fail the capacity checks even for the unmodified existing loads, which effectively requires the engineer to maintain two separate sets of values; one in WSD format for the existing piles, and the other in LSD format for the new piles. This again undermines credibility of the LSD paradigm in the eyes of the practitioner, who still struggles to understand why all this change was needed in the first place.

3 LSD CALIBRATION BY PROBABILISTIC CONSIDERATIONS

In the Becker paper [5], and typically in the theory of LSD method, the calibration of the resistance factors was done as follows. The load \( \bar{L} \) and resistance \( R \) were taken as log-normally distributed and the probability of equation \( P(R < \bar{L}) \) controlled by stipulating the target value of reliability index, \( \beta \). The reliability index is the number of standard deviations between the mean value of the distribution, in this case, of the parameter \( z = \ln R - \ln \bar{L} \), and the “failure limit”, such that the area of the tail of the distribution lying below this failure limit is sufficiently small. Normally, \( \beta \) is targeted to be between 3.0 and 3.5 for structures, which
corresponds to the probability of failure from $10^{-3}$ to just under $10^{-4}$. Becker proposed to use $\beta = 3.4$ for deep foundations in compression designed by “semi-empirical analysis using in-situ and laboratory test data”.

With some approximation, Becker [5] obtained the following equations for the “practical” ranges of coefficients of variation of both the loads and the resistances:

\[
\alpha = k_L \exp(\theta \beta v_L) \\
\phi = k_R \exp(-\theta \beta v_R)
\]

where $k_L$ and $k_R$ = “bias factors” or multiples between the “characteristic” and mean value of the load and the resistance, respectively; $v_L$ and $v_R$ = coefficients of variation of the load and the resistance; and $\theta$ = “separation” factor, given by the formula:

\[
\theta = \sqrt{1 + \left(\frac{v_R}{v_L}\right)^2} \\
= \frac{1 + \left(\frac{v_R}{v_L}\right)^2}{1 + \left(\frac{v_R}{v_L}\right)^2}
\]

The ratio $v_R/v_L$, according to Becker, is difficult to quantify because of the lack of data, but in the range of this parameter from 0.2 to 5 the quantity $\theta$ does not change appreciably. At the ends of this interval, it equals 0.85, whereas at $v_R/v_L = 1$ it reaches its minimum value, equal to 0.7. Constant value of $\theta = 0.75$ is recommended by Becker, which corresponds to $v_R/v_L = 0.5$ or 2. However, this may not be accurate as the range of variation of $\theta$ is small. For $k_R$, Becker recommends a constant value $k_R = 1.1$. The factor $k_L$ reportedly varies from 0.8 to 1; for this study, $k_L = 0.9$ is assumed.

Equations (4a) and (4b) still imply an interrelation between the load factor and the resistance factor, linked together by the common parameter $\beta$. Changing this parameter modifies both $\alpha$ and $\phi$. It is still possible to satisfy the “compatibility” Equation (3) by setting:

\[
\ln \left( \frac{GFS \times k_R}{k_L} \right)
\]

For the adopted values of parameters, $v_L + v_R = 0.44$ to match the Global Safety Factor of 2.5. Equation (5) caps the total variability of both the load and the resistance, which is consistent with the idea of the single “all perils” factor of safety.

Instead, Becker [5] used data provided by Barker et al. [12] who proposed $v_R = 0.2$ to 0.4 based on the tests of bridge piles. The upper end of this range was assumed for NBCC calibration (Becker admits that the resistance factors adopted for the AASHTO code (1998), which were also based on the Barker’s data, are “generally greater” than the ones he recommended for the NBCC based on the same data.) The assumption of $v_R = 0.4$ seems too conservative considering Barker’s data, noting that $v_R$ is closer to 0.2 – 0.3 for piles designed using soil shear strength values determined by laboratory tests or the more refined in-situ tests such as Cone Penetration Test, and only reach to higher ranges when based on crude testing procedures such as Standard Penetration Test (SPT). It is unusual in Canadian design practice to base the pile design on SPT alone, without some assessment of the soil shear strength.

More important, however, is the fact that the value for $v_R$ was assigned directly from pile
bearing capacity tests, without regard to variability of loads. This constitutes a fundamental difference between probabilistic calibration and calibration for WSD compatibility. The concept of GSF to cover “all perils” has made convenient use of the fact that “two bombs don’t fall in one pit”; in other words, it is unlikely that both the load and the resistance are at their extreme values at the same time. When calibrating based on the results of resistance tests alone, this beneficial effect is lost. It is seen from Equation (6) that assuming $v_R = 0.4$ leaves no room for variability of loads to still be compatible with the GSF value of 2.5.

On the “load” side of the equation, if we attempt to justify the uniform value of the reliability index $\beta = 3.4$ on the load factor value of $\alpha = 1.45$ as adopted in the Canadian code [3], Equation (4a) would give $v_L = 0.19$. The general ranges of variation of this parameter from studies by Allen, MacGregor and Meyerhof, cited by Becker [5], are from 0.07 to 0.3 and even greater, depending on the type of load (less for dead loads, more for live and environmental loads). Since we are considering buildings where live loads are substantially greater than dead loads, the $v_L$ should be closer to the upper end of the variation range. Because of the presence of exponential functions in Equations (4a) and (4b), the result of the calculation is extremely sensitive to the choice of the parameters $v_L$ and $v_R$, which can only be back-calculated from data for existing designs or evaluated on basis of a relatively small number of tests. However, all this number matching is only necessary to establish the same degree of reliability that has been already achieved by WSD. If so, then again, what is the purpose or advantage of the transition to LSD?

This intuitively obvious fact was confirmed numerically by modelling two log-normally distributed random variables – “load”, with mean value 200 kN and standard deviation 60 kN, and “resistance”, with mean value 500 kN and standard deviation 150 kN, which results in the coefficient of variation values $v_L = 0.3$ and $v_R = 0.3$. These values deviate somewhat from those chosen by Becker [5] for his calibration; however, they are in a reasonable and practical range. The ratio between the mean values was chosen to correspond with GSF = 2.5, which is known to be adequate by experience. 32,000 realizations of each of the variables were obtained using MS Excel random number generator.

As a result, the probability of underrunning the resistance $\phi R_n$ as per Equation (4b), where $\phi = 0.4$ and $R_n = \bar{R}/k_R = 500/1.1 = 450$ kN, was found to be 0.04%, which corresponds to the $\beta$ factor of 3.35 as targeted by Becker for the resistance side of the equation alone. However, the probability of exceeding the load $\alpha L_s$ as per Equation (4a), where $\alpha = 1.45$ and $L_s = \bar{L}/k_L = 200/0.9 = 220$ kN, was found to be 3.9%, which corresponds to the $\beta$ factor of 1.765. Note that using Equation (4a) and adopting $\alpha = 1.45$, $k_L = 0.9$, $\theta = 0.75$ and $v_L = 0.3$, we would obtain $\beta = 2.12$ and the probability of exceedance of only 1.7%. This discrepancy is likely caused by the approximate nature of the “separation” factor $\theta$ in the LSD calibration and the extreme sensitivity of the Equations (4a) and (4b). In any case, however, the $\beta$ factor for loads is a distant cry from the imposed target for resistances. Using a single value of $\beta$ for both the load and the resistance in this calculation, in the writer’s opinion, is invalid.

The “proper” way of formulating Equation (6) is as follows:

$$\beta_L v_L + \beta_R v_R = \frac{1}{\theta} \ln \left( GFS \times \frac{k_R}{k_L} \right)$$

(7)
This equation is closely satisfied by the adopted values of parameters matching the GSF of 2.5, whereas Becker’s chosen values \((v_L = 0.2, \; v_R = 0.4, \; \beta = 3.4, \; \theta = 0.75)\) lead to GSF = 3.8, which is consistent with the combination of \(\alpha\) and \(\phi\) he recommended. Does it mean that the WSD design procedure was inadequate and a more stringent procedure was required?

In [4], Becker provided “averaged” data for foundations’ probability of failure being \(10^{-2}\) to \(10^{-3}\) per year. This corresponds to the empirical \(\beta = 2.3\) to 3.1. The increase of the target \(\beta\) to 3.4 must indicate that the society is dissatisfied with the current state of safety of its pile foundations. However, the writer is unaware of any such widespread concern. Likely the adopted values of \(v_L\) and \(v_R\) are too high when considered together; but there is no mechanism for coordinating them in the method with separate load and resistance factors.

The correction of existing safety factors by new test data may be described as a procedure of Bayesian update. The prior probability is characterized by the conventional GSF; the resulting probability of failure has been calculated and graphed by Becker [4]. The likelihood function is the new probabilistic data from tests such as those described in [8, 9, 10] and others. The resulting posterior probability of failure is going to be close to the prior probability unless either the likelihood function deviates significantly from the existing data, or the amount of the experimental data becomes so massive as to compete with the tremendous number of foundations designed using the WSD and performing adequately.

4 CONCLUSIONS

- The paper discusses transition from Working Stress Design (WSD) using a single Global Safety Factor (GSF) to Limit States Design (LSD) using differentiated load and resistance factors in one specific aspect of foundations engineering – design of pile foundations for vertical compressive loads. Using this case as an example, the paper examines a more general question of probabilistic calibration of load and resistance factors in LSD and similar methods.

- Published data on “averaged” foundations’ probability of failure correspond to the reliability index \(\beta = 2.3\) to 3.1, which is substantially lower than the target value of 3.4 in the 2010 National Building Code of Canada [3]. Therefore, the new load and resistance factors have been effectively set to increase the level of design reliability (and therefore the cost) of pile foundations.

- The traditional design concept of WSD has served the construction industry well for decades. There are no established facts of it being inadequate as a method (although poor designs, as well as good designs, can be produced in this and any paradigm). In the author’s opinion, there is no reason to exceed the equivalent safety margins used in the traditional designs.

- It is shown analytically and confirmed numerically by Monte Carlo simulation that uniform level of reliability is not achieved in LSD method. The imposed reliability of resistance is generally much higher than the reliability of load calculation, which limits usefulness of these methods for direct probability-of-failure-based design.

- The true value of LSD is in the more orderly approach to evaluation of modes and probabilities of failure. As such, it should have been reserved to research and demonstration purposes only. However, since it has penetrated into the field of practical design, the probabilistic calibration should not withstand the customarily
established values of Global Safety Factors (Re: Equation (3)), unless a compelling body of evidence is assembled to prove the inadequacy of the traditional methods.

- The existing load factors for pile foundations embodied in [3, 6], in the opinion of the writer, need to be recalibrated to conform with the customarily adopted Global Safety Factors from historical WSD designs [7].

- A procedure of Bayesian update is proposed for development of improved code provisions based on new research data. Within its framework, code modification will only occur if the results of new tests deviate significantly from the existing data, or the amount of the new test data becomes comparable with the tremendous number of foundations designed under WSD and performing adequately.

REFERENCES


A GENERAL APPROACH FOR DECOUPLING RBDO PROBLEMS

ANDRE J. TORII¹, RAFAEL H. LOPEZ² AND LEANDRO F. FADEL MIGUEL³

¹Center for Optimization and Reliability (CORE), Department of Scientific Computing, Universidade Federal da Paraiba, Cidade Universitaria, Joao Pessoa, Brazil ajtorii@gmail.com

²Center for Optimization and Reliability (CORE), Department of Civil Engineering, Universidade Federal de Santa Catarina, Rua Joao Pio Duarte Silva, s/n, Florianopolis, Brazil rafael.holdorf@ufsc.br

³Center for Optimization and Reliability (CORE), Department of Civil Engineering, Universidade Federal de Santa Catarina, Rua Joao Pio Duarte Silva, s/n, Florianopolis, Brazil leandro.miguel@ufsc.br

Key words: Reliability based design optimization, shift vector, reliability accuracy, sequential optimization and reliability assessment

Abstract. In this paper, a novel approach for Reliability Based Design Optimization (RBDO) is presented. In this approach, we employ the concept of shifting vectors, originally developed in the context of the Sequential Optimization and Reliability Assessment (SORA). However, shifting vectors are found and updated based on a novel strategy. The resulting framework is able to use any technique for reliability analysis, such as Monte Carlo Simulation, among others. The main advantage of the approach is that accuracy of reliability analysis depend on the technique used, and thus can be improved. Several numerical examples are presented in order to highlight the advantages of the proposed method.

1 INTRODUCTION

It is widely acknowledged that deterministic optimization is not robust with respect to the uncertainties which affect engineering design. In deterministic optimization, potential failure modes of the designed engineering system are converted in design constraints, and uncertainty is addresses indirectly, by means of safety coefficients and conservative
assumptions. This approach is inherited from design through design codes, which is essentially non-optimal (since approximations are always conservative). On the other hand, a deterministic optimal design has, naturally, more failure modes designed against the limit. Hence, deterministic optimal designs are potentially less safe than their non-optimal counterparts. In this context, reliability-based design optimization (RBDO) has emerged as an alternative to model the safety-under-uncertainty part of the problem.

In RBDO, one usually looks for the minimization of some objective function involving material or manufacturing costs subject to constraints on failure probabilities. The RBDO formulation explicitly addresses the uncertainties affecting system performance and ensures that a minimum specified level of safety is maintained by the optimum design. Since design optimization is an iterative process, the failure probability of each constraint must be evaluated at each iteration. It may be accomplished by different approaches, for instance, simulations methods (e.g. Monte Carlos simulation (MCS), Importance Sampling), the first and second order reliability methods (FORM and SORM, respectively) and full characterization methods (e.g. polynomial chaos based approaches). The choice among these approaches is usually a ratio between accuracy and computational cost.

In the last two decades, the most employed method for the reliability analysis in RBDO problems was the FORM. Since it is an optimization procedure itself, the solution of the RBDO is a double-loop strategy, i.e. structural optimization and reliability analysis, which leads to high computational costs. To reduce the computational burden of the RBDO, several researchers focused on the decoupling of the structural optimization and the reliability analysis. These methods have been compared in several benchmark problems by different researchers [1, 2] and the Sequential Optimization and Reliability Assessment (SORA) [3] was one of the most promising techniques.

Despite of all these developments, the issue of accuracy remains an open question, due to the limitations of the FORM. It only provides the exact value of the probability of failure if the limit state function is linear and all the random variables are independent and normal, which is a rather limited case in practice. For this reason, some authors started to pursued the development of decoupling techniques which allow for the utilization of any reliability analysis method, such as Zou and Mahadevan [4]. In this context, we propose an approach for RBDO that allows any reliability analysis technique to be employed. The reliability and the optimization steps are decoupled by using a strategy based on shift vectors of the SORA method, which are found as to enforce the required reliability level. Also, the proposed approach does not require gradient information of the probabilistic constraints due to the use of a least squares approximation. The resulting algorithm is able to solve problems that simply cannot be addressed by RBDO techniques based on the FORM. It should be remarked that the proposed approach was not developed to be more computationally efficient than FORM based strategies such as SORA. The main goal of the proposed approach is to improve accuracy of the reliability analysis step. By allowing any reliability analysis technique to be used, the proposed approach broadly extends the range of application of RBDO.
This paper is organized as follows: Section 2 presents the formulation of the general RBDO problem. The concepts of the SORA method are detailed in Section 3 and the proposed RBDO approach is presented in Section 4. Several numerical examples are solved in Section 5. Finally, the main conclusions drew from this work are presented in Section 6.

2 THE RBDO PROBLEM

A RBDO problem can be stated as

Find: \( d \in R^n \) (1)

that

\[ \min f(d) \] (2)

subject to

\[ P_f = P(G(d, X) > 0) \leq P_t, \] (3)

where \( d \) is the design vector, \( f \) is the objective function to be minimized, \( G \) is a nonlinear limit state function, the vector \( X \in R^n \) is comprised by the random variables of the system under analysis and \( P(\cdot) \) is the probability of a given event occurring. Here, we assume the case of a single inequality constraint, but the results are extended for other cases later on this paper. In this problem, we search for the design vector \( d \) that minimizes the objective function subject to a probabilistic constraint. The probabilistic constraint from Eq. (3) actually states that the failure probability \( P_f \) of the original constraint being violated must be smaller than a target failure probability \( P_t \).

In RBDO, the constraint from Eq. (3) is frequently rewritten in terms of reliability indexes in order to avoid manipulation of very small numbers. In this case, the constraint can be rewritten as

\[ \beta \geq \beta_t \] (4)

where \( \beta \) is the reliability index related to the limit state function \( G \) and \( \beta_t \) is a target reliability index, which are defined respectively as

\[ \beta = \Phi^{-1}(1 - P_f) \quad \text{and} \quad \beta_t = \Phi^{-1}(1 - P_t), \] (5)

where \( \Phi^{-1} \) is the inverse cumulated density function of a standard normal random variable.

3 Sequential optimization and reliability assessment

In this section, we first present the basic concepts of SORA and then we present a different view of this method, which we use to developed the proposed approach.
3.1 A brief review of SORA

Here, only the basic concepts of SORA are presented, for a detail explanation the reader is referred to Du and Chen [3]. SORA is based on the strategy of serial single loops decoupling the optimization and the reliability analysis in RBDO. At each iteration of the method, the reliability analysis is only conducted after convergence of several loops of the optimization problem. The key concept of the method is to shift the boundaries of the violated equivalent deterministic constraints to the feasible direction based on the reliability information obtained in the previous cycle, which makes the reliability constraints improve progressively.

The proposed approach uses the concept of a shift vector in order to solve the RBDO problem, originally developed in the context of SORA. In this case, we assume that the original RBDO problem can be substituted by the approximate deterministic optimization problem

\[
\text{Find: } d \in \mathbb{R}^n \\
\text{that} \\
\min f(d) \\
\text{subject to} \\
g(d, \mu_X + s) \leq 0,
\]

where \( g \) is the equivalent deterministic constraint of \( G \), \( \mu_X \) is the vector comprised by the mean values of the random parameters \( X \), and \( s \) is a shift vector, which is defined as to ensure that satisfaction of the constraint from Eq. (8) enforces the satisfaction of the constraint from Eq. (3).

One of the main advantages of SORA is that it requires the solution of the deterministic optimization problem given by Eqs. (6)-(7). This fact simplifies the computational implementation of the approach, since existing deterministic optimization algorithms can be easily adapted. However, in SORA, the shift vector is built using the concept of the Hasofer-Lind reliability index, and consequently it may be inaccurate in several cases. In fact, SORA is able to reduce the computational effort of the RBDO problem, but it is not able to improve accuracy of the reliability analysis. In this case, the accuracy from the probabilistic point of view is entirely dominated by the accuracy of the FORM.

3.2 An alternative view of SORA

Here, we propose an alternative view of the use of shift vectors for RBDO. In this case, we note that a given shift vector \( s \) will ultimately result in a given optimum design vector \( d \) after the solution of the minimization problem from (6)-(8). This allows us to build a relation between \( s \) and \( d \) as follows:
which is given by solving the deterministic optimization problem from (6)-(8) for a given shift vector $s$. On the other hand, a given design vector $d$ results in a given failure probability $P_f$, thus defining the relation

$$R : d \rightarrow P_f,$$

which is given by reliability analysis using the design vector $d$. Combining Eqs. (9) and (10) we can build a composite relation

$$J = R \cdot O(s) : s \rightarrow P_f,$$

which is a functional relating shift vectors to failure probabilities. That is, we assume that every shift vector can be associated to a given failure probability, thus defining a functional. Of course this relation is defined implicitly by means of an optimization problem and a reliability analysis procedure, but this is not important for now.

We note that the relation from Eq. (11) can be written in terms of the reliability index as

$$J : s \rightarrow \beta,$$

where we assume that the reliability index is as defined in Eq. (5). This leads to the conclusion that the reliability index of the optimum solution depends on the shift vector used and write we $\beta(s)$.

The main idea is to find the shift vector $s^*$ that gives $P_f = P_t$. When this is accomplished, the optimum design vector $d^*$ that gives the desired reliability can be found from Eqs. (6)-(8) using $s = s^*$. We note that this is precisely the idea of SORA, where the Hasofer-Lind reliability index is used in the reliability analysis step and some more developments are made.

4 Proposed approach

This section presents the proposed approach to decouple the structural optimization and the reliability analysis, which allows for the utilization of any reliability analysis method.

4.1 Main idea

We tackle the RBDO problem taking the shift vector as the primary variable. The RBDO problem can then be stated as

$$\text{Find: } s$$

that
\[
\min \left[ \beta_t - \beta(s) \right]^2 \quad (14)
\]
subject to
\[
\beta_t - \beta(s) \leq 0. \quad (15)
\]

Once this problem is solved, we can recover the optimum design vector from Eqs. (6)-(8) using \( s = s^* \).

The constraint from Eq. (15) ensures that the reliability index of the design is higher than the target reliability index. Besides, the objective function from Eq. (14) tries to enforce \( \beta = \beta_t \). In most cases, the algorithm will actually make \( \beta = \beta_t \). However, we leave the problem as stated here in order to allow the reliability index to be higher than the target if needed. This is important in problems subject to more than one constraint, since in these cases it is common that the reliability index of only a few constraints match \( \beta_t \), i.e. only a few probabilistic constraints are active.

Obviously, it is not easy to solve the problem given by Eqs. (13)-(15). In order to evaluate \( \beta \), it is necessary to solve the optimization problem that defines Eq. (9) and to make the reliability analysis that defines Eq. (12). Thus, \( \beta(s) \) is defined implicitly. However, here we build a least squares approximation for the functional \( J \), which allows the efficient solution of the problem. Such an approximation is detailed in the next section.

### 4.2 Least Squares approximation for \( J \)

Instead of solving the problem given by Eqs. (13)-(15) directly, the problem can be solved for some approximation for \( J \). Here, we use linear approximations

\[
J(s) \approx J_{LS}(s) = a_0 + \sum_{i=1}^{n} [b_i s_i], \quad (16)
\]

and quadratic approximations

\[
J(s) \approx J_{LS}(s) = a_0 + \sum_{i=1}^{n} [b_i s_i] + \sum_{i=1}^{n} [c_i s_i^2], \quad (17)
\]

where \( a_0, b_i \) and \( c_i \) are coefficients found by Least Squares approximation.

We note that \( J(s) \) must be evaluated for at least \( n + 1 \) different shift vectors if a linear approximation is to be made. In the case of the quadratic approximation, \( 2n + 1 \) shift vectors are necessary. This least squares approximation may be viewed as a response surface approximation and for the interest reader details of the implementation may be found in Torii and Lopez [5]. The implementation of the proposed approach is given in the next section.
4.3 Algorithm

The problem from Eqs. (13)-(15) can be solved by the following steps.

First, since we do not know $J(s)$ explicitly, it is interesting to build some approximation for it. A Least Square approximation can be used for this purpose. We then generate a set of different shift vectors

$$ S = \{s_1, s_2, ..., s_m \} \quad (18) $$

and solve the deterministic optimization problem given by Eqs. (6)-(8) using $s = s_1, s_2, ..., s_m$.

For example, if the designer decides to start with a linear approximation, $m$ must be at least equal to $n + 1$. For each shift vector, we find a design vector $d = d_1, d_2, ..., d_m$. We then pursue $m$ reliability analyses using $d_1, d_2, ..., d_m$, resulting in the reliability indexes $\beta(s_1), \beta(s_2), ..., \beta(s_m)$, respectively.

We now use the approximation $J_{LS}$ in the place of $J$ in the problem from Eqs. (13)-(15) and find an approximate solution for $s$. This can be solved using any constrained minimization algorithm. Since $J_{LS}$ is an analytical and simple function, the computational effort required is low.

This gives a new shift vector that can be included in the set $S$ from Eq. (18). With this new shift vector, we can find a new design vector $d$ and evaluate the new reliability index. With this information, the approximation $J_{LS}$ can be rebuilt. The procedure is repeated until convergence is achieved.

The proposed algorithm can then be stated as:

1. Make $k = 0$ and generate $m = n+1$ shift vectors to build the set $S(k) = \{s_1, s_2, ..., s_m \}$, where $k$ is the iteration number;
2. For each shift vector from $S$, solve the optimization problem of Eqs. (6)-(8) and obtain the corresponding design vectors $\{d_1, d_2, ..., d_m \}$;
3. For each design vector, pursue the reliability analysis using the desired method, evaluating the reliability indexes $\{\beta(s_1), \beta(s_2), ..., \beta(s_m) \}$;
4. Build a Least Squares approximation $J_{LS}(s) \approx J(s)$ relating the shift vectors to the reliability indexes;
5. Solve the minimization problem from Eqs. (13)-(15) and find a new shift vector $s_k$;
6. For the new shift vector solve the optimization problem of Eqs. (6)-(8) and obtain the corresponding design vector $d_k$;
7. For the new design vector $d_k$ pursue the reliability analysis, finding the corresponding reliability index $\beta(s_k)$;
8. Check convergence: if converged, then stop, otherwise update the shift vector set as $S_{k+1} = \{s_1, s_2, ..., s_m, s_k\}$, make $k = k + 1$ and return to step (4).

The main goal of the proposed approach is to allow the reliability analysis to be made using any strategy. The algorithm requires the evaluation of the reliability index $\beta$ (or $P_f$), but there is no requirement on which method to use, as it may be seen in the description of the proposed method above. In the numerical analysis section, we employ the crude MCS and FORM, but any other method can be used. Accuracy of the results (from the probabilistic point of view) can be improved by using more accurate reliability methods. We note that most of the other RBDO strategies are not flexible in this part. In fact, the SORA can only be made assuming the FORM approximation, which is known to be very poor in some cases. We also note that the proposed approach solves the deterministic optimization problem given by Eqs. (6)-(7), as occurs in SORA. This simplifies the computational implementation of the approach, since existing deterministic optimization algorithms can be easily adapted.

5 NUMERICAL EXAMPLES

It should be noted that the optimization problem given by Eqs. (6)-(8) for a fixed $s$ can be solved using any deterministic optimization algorithm. The same holds for the problem given by Eqs. (13)-(15) solved for $J_{LS}$. In this section, a Sequential Quadratic Programming (SQP) is employed in both cases. All examples were solved in MATLAB using a machine with 2.5GHz processor and 2Gb RAM.

In the following examples we assume that a negative reliability index indicates that the failure probability is higher than 50%. Besides, when no failures are captured by MCS (i.e. the reliability index is very big but unknown) we write $\beta = \infty$.

5.1 Example 1: 10 bars truss optimization problem

We study here the RBDO of the truss structure presented in Fig. 1 (Yi and Cheng, 2008). The structure dimensions are taken as $b = h = 360$ in. The magnitudes of the two applied forces are independent lognormal random variables with LN(1.0E5, 5.0E3) lb. The yielding stress is a normal random variables with N(2.5E4, 5.0E3) psi. All these random variables are grouped into the random vector $X$. We take the cross sectional areas as design variables and treat them as deterministic variables. The total volume of the structure is to be minimized. Lower and upper bounds on the design variables are defined as $0.01$ in$^2$ and $25.0$ in$^2$. The initial design is taken as $d_i = 10.0$ in$^2$. The target reliability index is $\beta_t = 3.0$.

In this problem, we include several types of limit state functions. First, we define the allowable displacement limit state function as

$$G_1(d, X) = \max(|u(d, X)|) - 4.5,$$

where $u$ is vector of nodal displacements. The stress limit state functions are defined as
Figure 1: Ground structure composed by 10 bars

\[ G_{1+i}(d, X) = |\sigma_i(d, X)| - f_y, (i = 1, 2, \ldots, 10), \]  

(20)

where \(\sigma_i\) is the stress inside bar \(i\) and \(f_y\) is the yielding stress of the material.

The problem was first solved using SORA with all limit state functions from Eqs. (19)-(20). The solution found was \(d=(11.56, 7.64, 15.28, 0.01, 0.01, 0.01, 10.80, 8.27, 0.01, 10.80)\) in\(^2\), which is essentially the same found in literature [6].

In practice, it makes sense to consider that all limit state functions from Eqs. (19)-(20) must be respected at the same time and thus impose

\[ \overline{G}(d, X) = \max_{i=1,2,3,\ldots,11} G_i(d, X) \leq 0. \]  

(21)

This is the same as assuming that failure occurs if any of the two conditions are met: i) any stress constraint is violated, ii) the displacement constraint is violated. However, FORM may give a very poor approximation for the failure probability in this case.

If we subject the solution found by SORA to reliability analysis using MCS (\(ns = 1,000,000\)) and using the limit state function \(\overline{G}\), we obtain the actual reliability index of the structure as \(\beta = 2.7178\), which is lower than the target. We also note that SORA does not converge if the limit state function \(\overline{G}\) is used instead of the set \(G_i\). However, if SORA was able to converge, the solution would likely be inaccurate from the probabilistic point of view, since the Hasofer-Lind reliability index is a poor approximation in the case of the limit state function \(\overline{G}\).

We then applied the proposed approach for the problem using the limit state function \(\overline{G}\) and making linear approximations for \(\beta\). Since FORM gives a poor approximation in this case, reliability analysis is made using MCS with \(ns = 1,000,000\). The reliability
The solution found is presented in Table 1. The solution found is $d=(14.69, 7.36, 14.72, 0.01, 0.01, 0.01, 10.40, 10.36, 0.01, 10.40)$

Table 1: Example 1: Results obtained with proposed approach

<table>
<thead>
<tr>
<th>Iteration</th>
<th>$\beta$ ($ns = 1,000,000$)</th>
<th>Approximation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-1.1817</td>
<td>-</td>
</tr>
<tr>
<td>0</td>
<td>-1.2632</td>
<td>-</td>
</tr>
<tr>
<td>0</td>
<td>-0.3048</td>
<td>-</td>
</tr>
<tr>
<td>0</td>
<td>-0.4535</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>1.8368</td>
<td>Linear</td>
</tr>
<tr>
<td>2</td>
<td>-0.2148</td>
<td>Linear</td>
</tr>
<tr>
<td>3</td>
<td>7.0345</td>
<td>Linear</td>
</tr>
<tr>
<td>4</td>
<td>0.2548</td>
<td>Linear</td>
</tr>
<tr>
<td>5</td>
<td>0.3784</td>
<td>Linear</td>
</tr>
<tr>
<td>6</td>
<td>2.0790</td>
<td>Linear</td>
</tr>
<tr>
<td>7</td>
<td>2.8336</td>
<td>Linear</td>
</tr>
<tr>
<td>8</td>
<td>2.9635</td>
<td>Linear</td>
</tr>
<tr>
<td>9</td>
<td>2.9833</td>
<td>Linear</td>
</tr>
</tbody>
</table>

The solution presents a reliability index that is very close to the target. This puts in evidence that the problem solved for the limit state function $G$ is somewhat different from the one solved for the limit state functions $G_i$. We also note that FORM based methods will not be able to obtain an accurate solution for $G$, since FORM is inaccurate in this case.

5.2 Example 2: Problem with discrete random parameter

The last example is given by

FIND: $d = (\mu_{X_1}, \mu_{X_2})$ (22)

that

$$\min f(d) = \left(3\mu_{X_1}^2 + 2\mu_{X_2}\right)^2$$ (23)

subject to the constraint from Eq. (4) and box constraints

$$0 \leq d_i \leq 10(i = 1, 2),$$ (24)

with limit state functions

$$G_1(d, X) = (10 + 5X_3) - (X_1^2 + 2X_2^2)$$ (25)
\[ G_2(d, X) = (1 + X_3) - X_1. \] (26)

We note that \( X_3 \) is not a design variable, but just a random parameter of the problem.

The random variable \( X_1 \) has normal distribution with standard deviation equal to \( 0.2\mu_{X_1} \). The random variable \( X_2 \) has uniform distribution with lower and upper bounds given by \([0.5\mu_{X_2}, 1.5\mu_{X_2}]\). The random parameter \( X_3 \) is a discrete random variable with a probability of 0.7 of assuming the value 1.0 and a probability of 0.3 of assuming the value 0.0. The target reliability index is \( \beta_t = 2.0 \). We note that the discrete random parameter \( X_3 \) cannot be treated directly by FORM based approaches.

The proposed approach was applied using MCS (\( ns = 1e7 \)) for the reliability analysis. The results are presented in Table 2, where \( \beta_1 \) and \( \beta_2 \) are the reliability indexes for the two limit state functions. We note that the algorithm converged for a solution whose reliability indexes are close to the target.

<table>
<thead>
<tr>
<th>Iteration</th>
<th>((x_1, x_2))</th>
<th>((\beta_1, \beta_2))</th>
<th>Approximation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>(2.4000, 2.3455)</td>
<td>(0.4779, 1.0715)</td>
<td>-</td>
</tr>
<tr>
<td>0</td>
<td>(2.0000, 2.7452)</td>
<td>(0.5862, 0.3806)</td>
<td>-</td>
</tr>
<tr>
<td>0</td>
<td>(1.8000, 2.3196)</td>
<td>(0.0986, -0.0034)</td>
<td>-</td>
</tr>
<tr>
<td>0</td>
<td>(2.0000, 2.3453)</td>
<td>(0.2277, 0.3806)</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>(2.8856, 3.6051)</td>
<td>(1.8729, 1.7076)</td>
<td>Linear</td>
</tr>
<tr>
<td>2</td>
<td>(3.0284, 3.6685)</td>
<td>(2.0387, 1.8597)</td>
<td>Linear</td>
</tr>
<tr>
<td>3</td>
<td>(3.0284, 3.6685)</td>
<td>(1.8729, 1.7076)</td>
<td>Linear</td>
</tr>
<tr>
<td>4</td>
<td>(3.0812, 3.6191)</td>
<td>(2.0347, 1.9127)</td>
<td>Quadratic</td>
</tr>
</tbody>
</table>

This example illustrates the flexibility of the propose approach in dealing with complex situations. In particular, we note that FORM cannot be applied directly for discrete random variables. The proposed approach, on the other hand, is able to tackle the problem directly if MCS is used for the reliability analysis step.

6 Conclusions

In this work, we presented an approach for RBDO that allows any reliability analysis technique to be used. The reliability and the optimization steps are decoupled by using a strategy based on shift vectors, which are found as to enforce the required reliability level. The resulting algorithm is able to solve problems that simply could not be addressed by RBDO techniques based on FORM. Besides, as occurs in SORA, the approach requires the solution of a deterministic optimization problem, what simplifies its computational implementation.

The proposed approach was not developed to be more computationally efficient than FORM based strategies such as SORA or similar methods. The main goal of the proposed
approach is to improve accuracy of the reliability analysis step. By allowing any reliability analysis technique to be used, the proposed approach broadly extends the range of application of RBDO. FORM is known to be inaccurate in several cases of practical interest, what impedes the application of FORM-based RBDO strategies such as SORA. The accuracy of the reliability analysis in the proposed approach, on the other hand, can be improved by using more accurate techniques.

Even though FORM may be employed in the reliability analysis step of the proposed method, it makes more sense to use SORA, or other efficient technique when FORM is accurate. However, the proposed method allows the utilization of any reliability analysis method, for example, from the crude MCS to full characterization approaches. This allows the solution of much more complex problems, such as the one including a discrete random parameter presented here. It is also possible to solve problems with highly nonlinear limit state functions, such as the ones defined as being the maximum among a set of functions.

7 ACKNOWLEDGEMENTS

The authors acknowledge the financial support of the Brazilian agencies CNPq and CAPES.

REFERENCES


EFFECT OF EXPECTED CONSEQUENCES OF FAILURE ON OPTIMAL STRUCTURAL DESIGN

THAÍS G. PEDROSA¹ AND ANDRÉ T. BECK²

¹Graduate Student, Dept. of Structural Engineering, Univ. of São Paulo, São Carlos, SP, Brazil
Av. Trabalhador são-carlense, nº 400, Centro. CEP 13566-590 São Carlos, SP, Brazil
thaisgp07@usp.br

²Assoc. Prof., Dept. of Structural Engineering, Univ. of São Paulo, São Carlos, SP, Brazil
Av. Trabalhador são-carlense, nº 400, Centro. CEP 13566-590 São Carlos, SP, Brazil
atbeck@sc.usp.br

Key words: Structural Reliability, Structural Optimization, Non-linear Analysis.

Abstract. Defying failure is the primary challenge of the structural engineer. It sounds paradoxical, but in order to achieve a successful design, the structural engineer must think about and account for all possible failure modes of a structure. This is no different in structural optimization. Hence, in structural optimization one also has to consider the expected costs of failure. In structural engineering design, economy and safety are apparently conflicting goals. However, when expected costs of failure are considered, it becomes evident that investments in safety are necessary in order to avoid paying for failure. The optimum point of compromise can be found by a risk optimization, where the objective function includes all costs over the life-cycle of the structure: construction, operation, inspection, maintenance, disposal, and the expected costs of failure. The latter are an undeletable remainder of the failure modes that the structure needs to be designed against. This paper addresses the optimization of simple structural systems, considering the balance between competing failure modes such as yielding (squashing), buckling and snap-through. The study shows how different failure modes, associated to different costs of failure, lead to different optimal designs. A plane truss structure is studied as application example. The shape (nodal positions) and member size are considered as design variables. Results show that different optimal designs are obtained when the balance between competing failure modes is changed.

1 INTRODUCTION

Structural optimization involves seeking for the (optimal) shape and size of a structure, which allows it to fulfill its intended function with minimal resources use. It is today widely known that deterministic design optimization is not robust w.r.t. uncertainties in structural loads and material strength. In deterministic design optimization, uncertainties are taken into account in a subjective and indirect way, by means of partial safety factors specified in design codes. As a consequence, deterministic optimal solutions may lead to reduced reliability levels [1]. As an example, consider a fully stressed design where all material points are design against the limit: such a structure is more likely to collapse, since there are no alternative load paths.
Thaís G. Pedrosa and André T. Beck

Structural optimization should to be robust with respect to load and resistance uncertainties. This perception has led to development of different methodologies:

1. Robust optimization [2,3];
2. Fuzzy optimization or fuzzy programming [4];
3. Reliability-based design optimization or RBDO [5].

Typical objectives in robust optimization are maximizing mean system performance and minimizing the variance. In robust optimization, uncertainties are represented as random variables or stochastic processes. Otherwise, in fuzzy optimization, uncertainties are represented by fuzzy numbers. While robust optimization has a probabilistic approach, the fuzzy optimization presents a possibilistic approach. In general, the amount of information available determines the appropriate approach [6].

RBDO seeks minimization of objective function involving material volume, manufacture cost or structural system performance. Uncertainty in loading and strength is addressed objectively and modeled probabilistically. Failure probabilities are normally used design constraints.

Consideration of uncertainties in structural design optimization is not a novelty. However, a literature review shows that in most classical RBDO articles [7,8,5,9,10,11,12,13] failure consequences are ignored or discarded.

Objective functions in robust optimization and RBDO frequently involve costs, but such costs relate to material volume or manufacturing costs. A cost term often overlooked in classical RBDO problems is the expected failure cost. Expected failure cost is the product between failure costs by failure probabilities. This definition complies with the definition of risk: expected cost of failure can be understood as the risk associated to each failure mode of the structure. The total expected cost is the sum of expected failure costs for each possible failure mode, in addition to initial construction costs, maintenance and operation costs. The total expect cost of a structural system is directly affected by the risk offered to users, employees, the general public and the environment.

In this study, the risk optimization formulation is employed [14,1,15]. In this formulation, expect costs of failure are included in the objective function. The optimization problem becomes unconstrained.

2 RISK OPTIMIZATION

Consider $X$ and $d$ two parameter vectors of a structural system. Vector $X$ represents all random variables of the problem such as dimensions, resistance properties of materials and structural members and loads. Some of these parameters are by nature random variables; others cannot be deterministically defined due to several uncertainty sources. Vector $d$ contains the system design variables or optimization variables, such as partial safety factors, dimensions, nodal coordinates and even some (deterministic) parameters of the probability distribution models of random vector $X$.

The existence of uncertainties implies risks, that is, the possibility of undesired structural responses. The boundary between desired and undesired structural responses is formulated in limit state equations $g(X,d) = 0$ such that:
\[ \Omega_f(d) = \{ x | g(X, d) \leq 0 \} \] (1)
\[ \Omega_s(d) = \{ x | g(X, d) > 0 \} \] (2)

where \( \Omega_f(d) \) and \( \Omega_s(d) \) are the failure and survival domains, respectively.

Each limit state function describes a possible failure mode in terms of serviceability or ultimate capacity. The probability of an undesirable response, or failure probability, is given by:

\[ P_f(d) = P[X \in \Omega_f(d)] = \int_{\Omega_f(d)} f_X(x) dx \] (3)

where \( f_X(x) \) represents the joint probability density function of random vector \( X \) and \( P[..] \) represents probability. The failure probability for each limit state function, as well as for system failure, is calculated using techniques of structural reliability, such as FORM (First Order Reliability Method), SORM (Second Order Reliability Method) or Monte Carlo simulation [16,17].

2.1 Quantifying failure consequences

The total expected cost \( C_{te} \) of a structural system in a risk optimization problem is given by the sum of the following costs:

1. \( C_{\text{initial}} \): initial or construction cost;
2. \( C_{\text{operation}} \): operation cost;
3. \( C_{\text{maint.}} \): Cost of inspections, maintenance, repair and replacement;
4. \( C_{\text{discard}} \): cost of discard;
5. \( C_{\text{ef}} \): expected cost of failure.

The expected cost of failure is the product between failure cost by the respective failure probability [1]:

\[ C_{\text{ef}}(d) = C_f(d)P_f(d) \] (4)

Failure costs include repair or replacement costs for damaged components, complete system rebuild, compensation costs due to accidental failure, and others. The failure probability is evaluated by structural reliability theory, following Eq. (3). To each failure mode there is an associated expected cost of failure term, hence:

\[ C_{te}(d) = C_{\text{initial}}(d) + C_{\text{operation}}(d) + C_{\text{maint.}}(d) + C_{\text{discard}}(d) + \sum_{k=1}^{\text{fail.modes}} C_f(d)P_f(d) \] (5)

The optimal amount of resources to invest in safety of the built structure, and in inspection and maintenance policies, is the one that leads to the minimum expected total cost. This amount can be found by solving the following risk optimization problem:

\[ d^* = \arg \min [C_{te}(d) : d \in S] \] (6)
where $S = \{d_{\min} \leq d \leq d_{\max}\}$ is the feasible design space and $d_{\min}$ and $d_{\max}$ are the lower and upper limits of the design variables. In this formulation, the reliability constraints are incorporated in the objective function.

The above formulation allows one to solve the safety-economy tradeoff in the design and operation of a given structure. Moreover, it allows one to find the optimal point of balance between competing failure modes, which is particularly important when significantly different failure costs are associated to different failure modes [18]. The risk optimization formulation allows finding the appropriate the appropriate safety margins w.r.t. each failure mode.

3 THE PROBLEM

The problem addressed in this paper is a simple two-bar truss, illustrated in Figure 1. Geometrical and material non-linearities are considered. The solution is formulated in analytical form.

In the initial configuration, the two bars form an angle $\theta_0$ with the imaginary horizontal line connecting the supports. Vertical force $P$ at node $B$ causes the vertical displacement $u$. The problem is symmetric w.r.t. the vertical line passing through $B$ and $B'$. The half-span is $b$ and height is $h$. The solution is parameterized w.r.t. $h/2b$.

By equilibrium condition, the relation between applied force and normal force in the bars is given by:

$$P(u) = N \sin \theta = \frac{N(h + u)}{L}$$

where $\theta$ is the bar angle in displaced position.

A non-linear elastic material is considered, with Green deformation:

$$E(u) = E_0(1 - \eta \varepsilon_G(u))$$

$$\varepsilon_G(u) = \frac{1}{2} \left( \frac{2hu + u^2}{L_0^2} \right)$$

By means of Hooke’s Law, the normal force is given as product of material elasticity constant by deformation:
Equilibrium paths for the structure, with geometrical and material nonlinearities, are shown in Figure 2, for $h/2b = 0.09$ and different material constants ($\eta$). Effect of pure geometrical nonlinearity can be observed for $\eta \to 0$. Note that with large deflections, as physical nonlinearity increases (larger $|\eta|$), more instability regions are observable, and limit loads are reduced.

![Figure 2: Equilibrium paths for different elastic nonlinearity constants for initial configuration ($h/2b = 0.09$).](image)

3.1 Variables and limit states functions

The bars are assumed hollow circular, with external diameter $d$ and thickness $t$. The vector of design variables is given by initial configuration and cross-section area: $\mathbf{d} = \{b, h, d, t\}$.

Three random variables are considered: applied load $P_a$, deformation modulus $E_0$ and yielding stress $\sigma_y$. The random variables follow Gaussian distributions with parameters $P_a \sim N(50,5)kN$, $E_0 \sim LN(20500,1025)MPa$ and $\sigma_y \sim LN(50,2.5)MPa$. Hence, random vector $\mathbf{X}$ is: $\mathbf{X} = \{P_a, E_0, \sigma_y\}$.

Three failures modes are considered:

6. Snap-through or limit point instability;
7. Buckling or bifurcation instability;
8. Yielding or squashing;

To formulate a solution, the final displacement, which delimits the equilibrium path extension, is first defined. This point is calculated by iterative Newton Raphson or by looking for a root of $P(u) - P_a$. The result is a displacement $u_f$ given as a function of $(P_a, E_0, \mathbf{d}, \eta)$.

Snap-through failure is defined by the limit point, which indicates the beginning of an unstable equilibrium. This point $u_{lim}$ is obtained by nullifying the tangent stiffness matrix. The limit displacement is function of $(h, b, \eta)$. The limit state function for snap-through failure is written as:

$$g_1(\mathbf{X}, \mathbf{d}, \eta) = P_{lim}(\mathbf{X}, \mathbf{d}, \eta) - P_a$$
where \( P_{\text{lim}} \) is the load corresponding to limit displacement \( u_{\text{lim}} \), and \( P_a \) is the applied load.

The second failure mode is bifurcation instability or buckling. The Euler relation (Eq. 12) that describes buckling is given by:

\[
N_{cr}(u) = \frac{\pi^2 E I}{L_0^2} = \frac{\pi^2 E_0 l (1 - \eta \epsilon_G(u))}{L_0^2}
\]

where \( N_{cr}(u) \) is the critical normal force. For bi-articulated bars, buckling length is equal to bar length. The limit state equation for buckling is:

\[
g_2(X, d, \eta) = N_{cr}(u_i) - N(u_i)
\]

where \( N(u_i) \) is the acting normal force, given in Eq. (10). During the load-displacement equilibrium path for the structure, Eq. (13) presents three local minima. The points of minima (\( u_i \) in Eq. (13)) are found by taking the derivative of Eq. (13) w.r.t. displacement \( u \) and equating it to zero. Failure due to buckling occurs at one of these points of minima or at the final displacement (\( u_f \)). The point of minima where buckling failure occurs changes according to the initial configuration (\( h/2b \)).

The third failure mode is compressive squashing or tensile yielding of the cross-section. The limit state function in Eq. (14) is given in terms of the yielding tension \( \sigma_y \), compared for both tensile and compression normal forces:

\[
g_3(X, d, \eta) = \min \left[ \left( \sigma_y - S_0 - \max_u \left[ |N_{\text{comp}}(u_j)| \right] \right) \right]
\]

where \( N_{\text{comp}}(u_j) \) and \( N_{\text{tensile}}(u_j) \) are evaluated from the load-displacement curve, by taking the derivative of normal force w.r.t. displacements \( (dN(u)/du) \), and by considering the final displacement. These displacements \( u_i \) in Eq. (14) are illustrated in Figure 3, which shows normal force versus displacement curves. One observes in Figure 3 the existence of two points of maximum compression, one point of maximum tensile force and the final displacement (black dot). The correct evaluation of the maximum normal (compression or tensile) force is done by a displacement analysis, which identifies the displacements that belong to the equilibrium path, according to final displacement \( u_f \). Figure 3 also shows the displacements \( u_i \) in Eq. (13). The difference between \( u_i \) and \( u_j \) is shown by the variation of minimum point in \( N(u) \) and \( N_{cr}(u) - N(u) \) curve.

![Comparison between acting and critical normal force.](figure3.png)
3.2 Initial and Failure Costs

The initial cost is given by material volume \((V = 2 L_0 S_0 \text{ cm}^3)\), multiplied by specific mass \(\rho (\text{g/cm}^3)\) and unit mass price (1000 $/kg):

\[
C_{\text{initial}}(d) = 2000 \rho c_{\text{unit}} L_0 S_0 = C_{\text{ref}} L_0 S_0
\]

where \(C_{\text{ref}}\) is a reference value for costs, \(L_0\) is given in \(cm\) and the area in \(cm^3\). The initial cost is, of course, function of the design variables.

It is well known that different failure modes are associated to different costs. Fragile or sudden failure, which occurs without warning, has more serious consequences than ductile failure. Service failures can lead to temporary suspension of use, which generally represents smaller costs than collapse (ultimate failure).

The simple two-bar structure considered herein could have many different applications; hence the costs of failure are immaterial. In order to avoid limiting the study to one possible application, the costs of failure are given generically in proportion to the initial cost.

Total expected cost can be written as:

\[
C_{\text{te}}(d) = C_{\text{initial}}(1 + A P_f \text{snap} + B P_f \text{buckling} + C P_f \text{tension})
\]

where \(A, B, C\), are failure cost factors that depend on the failure mode. Different failure cost scenarios are considered in the sequence.

4 RESULTS

In a first analysis, failure probabilities are evaluated by changing initial truss height, keeping other design variables fixed at \(d = 3.35 \text{ cm}, t = 3.4 \text{ mm}, b = 50 \text{ cm}\) and \(\eta = -100\). This study helps to understand the different failure mode behavior when the angle \(\theta_0\) is changed.

In Figure 4(a), one observes that the transition to snap-through failure is continuous, whereas transition to buckling failure is discrete. The yielding failure can be explained with aid of Figure 4(b), where each load-displacement curve corresponds to a configuration ratio in Figure 4(a). For large \(h/2b\) (\(h/2b > 0.95\)), normal load is compressive and failure may occur due to squashing (compressive yielding). For \(h/2b < 0.08\), snap-trough failure is highly likely,
and in this case there is a reversal in the normal load. Because of this reversal, there is a probability that the bars fail also due to tensile yielding.

Figure 5(a) shows results for a parametric analysis, where different costs of failure are considered. Figure 5(a) shows the objective function (\( C_{te} \)) in terms of truss height (above), diameter (center) and thickness (below). For each plot, the “remaining” design variables are fixed at: \( d = 3.35 \text{ cm}, t = 3.4 \text{ mm}, b = 50 \text{ cm} \) and \( h = 10.6 \text{ cm} \), which is the optimal height for the corresponding configurations. Total expected cost for this “reference” configuration, with \( A = B = C = 1 \), is \( C_{te}/C_{ref} = 0.1645 \). Reliability indexes for this configuration are: \( \beta_{snap} = 3.9 \), \( \beta_{buckling} = \infty \) and \( \beta_{yield} = 3.18 \).

In Figure 5(a), results are also shown for other cost configurations, were each cost term increased at once to ten units (10). Since buckling is not relevant for this configuration, one observes that changing the cost of failure for buckling does not change the objective function nor optimal values of design variables. However, increasing the cost of failure for snap-through and yielding affects the objective function, increasing optimal values of design variables \( h, d \) and \( t \). The change is not dramatic, but it shows that a different equilibrium point is obtained between the competing failure modes. Observed changes in design variables are of 3.77% for \( h \), and 4.76% for \( d \) and 6.45% for \( t \).

Figure 6(b) shows the objective function for other combinations of failure costs, in terms external diameter. The variation in total expected costs for the same scenarios is presented in Table 1.

### Table 1: Percentage change in design parameters for different cost configurations.

<table>
<thead>
<tr>
<th>( A )</th>
<th>( B )</th>
<th>( C )</th>
<th>( h/2b )</th>
<th>%</th>
<th>( d )</th>
<th>%</th>
<th>( t )</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.106</td>
<td>-</td>
<td>3.15</td>
<td>-</td>
<td>0.31</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>1</td>
<td>0.106</td>
<td>0</td>
<td>3.10</td>
<td>-1.59</td>
<td>0.31</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>1</td>
<td>0.106</td>
<td>0</td>
<td>3.20</td>
<td>1.59</td>
<td>0.32</td>
<td>3.23</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>10</td>
<td>0.110</td>
<td>3.77</td>
<td>3.30</td>
<td>4.76</td>
<td>0.33</td>
<td>6.45</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>20</td>
<td>0.112</td>
<td>5.66</td>
<td>3.35</td>
<td>6.35</td>
<td>0.34</td>
<td>9.68</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
<td>5</td>
<td>0.110</td>
<td>3.77</td>
<td>3.25</td>
<td>3.25</td>
<td>0.33</td>
<td>6.45</td>
</tr>
<tr>
<td>30</td>
<td>30</td>
<td>30</td>
<td>0.112</td>
<td>5.66</td>
<td>3.35</td>
<td>6.35</td>
<td>0.34</td>
<td>9.68</td>
</tr>
</tbody>
</table>
5 CONCLUSIONS

This paper addressed the optimal design of a non-linear truss structure, considering the effects of expected costs of failure. Physical and geometrical non-linearities were considered. Failure by yielding, buckling and snap-through were considered. Problem selection and its solutions targeted the competition between failure modes, when costs of failure are different. The differences between the optimal designs found herein were subtle, but they demonstrated how the competition between failure modes can affect optimal structural design.

REFERENCES


MECHANICAL AND RELIABILITY ANALYSES OF REINFORCED CONCRETE STRUCTURES SUBJECTED TO CORROSIVE EFFECTS CAUSED BY CHLORIDE IONS PENETRATION

GIOVANNI P. PELLIZZER¹, EDSON D. LEONEL² AND CAIO G. NOGUEIRA³

¹ MsC Structural Engineering, University of São Paulo. School of Engineering of São Carlos, Department of Structural Engineering. Avenida Trabalhador São Carlense, 400, 13566-590, São Carlos-SP, Brazil. gpais@usp.br

² Assistant Professor, University of São Paulo. School of Engineering of São Carlos, Department of Structural Engineering. Avenida Trabalhador São Carlense, 400, 13566-590, São Carlos-SP, Brazil. edleonel@sc.usp.br

³ Assistant Professor, São Paulo State University. Department of Civil and Environmental Engineering. Avenida Engenheiro Luiz Edmundo Carrijo Coube, 17033-360, Bauru-SP, Brazil. cgnogueira@feb.unesp.br

Key words: Mazars’ damage model; Fick’s second law; Reinforcements’ corrosion; Progressive collapse.

Abstract. Structural durability is an important design criterion, which must be assessed for every type of structure. In this regard, especial attention must be addressed to the durability of reinforced concrete (RC) structures. When RC structures are located in aggressive environments, its durability is strongly reduced by physical/chemical/mechanical processes that trigger the corrosion of reinforcements. Among these processes, the diffusion of chlorides is recognized as one of major responsible of corrosion phenomenon start. Reinforcements’ depassivation occurs when the chloride concentration at the reinforcement surface reaches a threshold value. Consequently, geometrical structural shape and mechanical proprieties of constituent steel are modified. The present work presents a formulation based on the Finite Element Method for the mechanical analysis of reinforced concrete structures subjected to corrosive effects due to chloride penetration. Mazars’s damage model and elastoplastic positive isotropic hardening approach are used to simulate physical nonlinearities of concrete and steel, respectively. Fick’s second law is used to model the chloride diffusion inside the structure and empirical laws are applied to represent the corrosion rate of reinforcements’ steel along time. The corrosion phenomena, as well as the mechanical proprieties of reinforced concrete, have a high level of randomness. Therefore, this problem is only properly addressed in the probabilistic context. Then, a polynomial response surface meta-model is built from the original mechanical model in order to represent the limit state equations. Monte Carlo’s Simulation Method is applied on the obtained limit state equation to determinate the failure probability. Hyper-static structures are analysed in order to determine the probability
of structural system failure, which is the main contribution of this study. Beams, localized in a saturated chloride environment, subjected to self-weigh and accidental loads are considered. The obtained results indicate that the progressive structural collapse due to corrosion is considerable different from the classical scenario when this chemical effect is not accounted.

1 INTRODUCTION

Concrete is the main construction type used in the world. Low relative cost, facility to obtain its components, durability and adaptability to various forms are some advantages of this material. However, every structure has a limited life time, affected by materials quality, design precautions, execution precision and especially by climate conditions.

Steel reinforcement corrosion is the most common pathology of concrete structures. Corrosion can be defined as the destructive interaction between a material and the environment. Chloride diffusion is one of the most important factors of the corrosion process [1-3]. The annual cost of corrosion worldwide is estimated to exceed $1.8 trillion, which translates 3-4% of the gross domestic product of industrialized countries [4].

The classical model of the deterioration process associated with reinforcement corrosion [5] defines two periods: initiation and propagation:

- Initiation: time to start steel reinforcement corrosion. Chloride induced corrosion is initiated jointly with reduced concrete cover, low quality concrete, poor compaction and curing. Chlorides ions may diffuse through the concrete cover, and corrosion is initiated when the chloride concentration reaches a threshold value. Corrosion also may be initiated if the structure has crack widths sufficiently large to allow the direct ingress of chlorides and other corrosion reaction components (oxygen, moisture, etc.).

- Propagation: time after the chloride concentration reaches a threshold value and the steel reinforcement integrity is altered. The steel mechanical capacity is reduced, as well as the cross section area of the bars. The expansive products of corrosion reaction produces internal tensile stresses that cause internal microcracking and, in advanced stages, external longitudinal cracking and eventually spalling. Other effect is the influence in the steel/concrete adherence.

However, as well as the mechanical proprieties of reinforced concrete, the parameters that influence the corrosion process have a high degree of randomness. The corrosive phenomenon has many associated uncertainties. Thus, a probabilistic approach may represent more consistently the corrosive process.

In this study, a formulation based on the Finite Element Method is presented in order to analyze reinforced concrete structures subjected to corrosive effects due to chloride penetration. Mazars’s damage model is used to simulate the behavior of concrete. The elastoplastic positive isotropic hardening approach is used to simulate the behavior of steel and an expression that associate the residual yield strength of steel as a function of corrosion amount is adopted. Fick’s second law is used to model the chloride diffusion inside the structure and empirical laws are applied to represent the corrosion rate of reinforcements’ steel along time.

Most of the available studies (related to corrosion, reliability analysis or both) are based on an isolated reinforced concrete element, generally, a beam element. It is important to note that the response of an isolated reinforced concrete element maybe not the same that of a complete
reinforced concrete structure [6]. This work is also incorporated in this context, dealing with
the analysis of hyper-static structures.

2 MECHANICAL MODEL

2.1 Damage model for concrete

The nonlinear mechanical behaviour of concrete originates from the crack growth along
the concrete mass. Damage models are particularly appropriate to represent this phenomenon,
because this approach allows the consistent representation of material degradation by
penalizing the material stiffness as a function of strain increase. In the present work, the
Mazars’ damage model [7] was adopted, which is based on the following hypotheses: damage
is an isotropic and scalar variable, the residual strains are totally neglected and damage occurs
only by elongation strains.

The state of elongation at a given point is represented by the equivalent strain, as follows:

\[ \varepsilon = \sqrt{(\varepsilon_1^+)^2 + (\varepsilon_2^+)^2 + (\varepsilon_3^+)^2} \]  (1)

in which \( (\varepsilon_i^+) \) corresponds to the positive components of the main strain tensor.

The criterion used to verify the material mechanical integrity is defined by:

\[ f(\varepsilon, D) = \varepsilon - S(D) \leq 0 \]  (2)

in which \( D \) corresponds to the damage scalar variable and \( S(D) \) represents the limit strain value as a function of the damage state. The damaging process starts when the equivalent strain at a given point is greater than the normal strain value of concrete tensile strength \( \varepsilon_{d0} \).

At the start of the concrete damaging process \( f(\varepsilon, D) = 0 \) and \( S(D) = \varepsilon_{d0} \). For the following steps, \( S(D) \) is updated considering the value of \( \varepsilon \) determined on the last converged load step, which accounts for the updated damage state.

Due to the non-symmetric mechanical behaviour of concrete when subjected to tensile and compressive stresses, the damage variable is composed by the sum of two independent parts: tensile portion, \( D_T \), and compressive portion, \( D_C \). Each of these portions indicates tensile and compressive contributions to the local strain state. These portions are obtained as a function of the equivalent strain and the internal parameters of the damage model, which are defined by:

\[ D_T = 1 - \frac{\varepsilon_{d0}(1 - A_T)}{\varepsilon} - \frac{A_T}{e^{[B_T(\varepsilon - \varepsilon_{d0})]}}; \quad D_C = 1 - \frac{\varepsilon_{d0}(1 - A_C)}{\varepsilon} - \frac{A_C}{e^{[B_C(\varepsilon - \varepsilon_{d0})]}} \]  (3)

in which \( A_T, B_T, A_C, B_C \) are the internal parameters of Mazars’ damage criterion, experimentally calibrated. The final value of the damage variable is composed as:

\[ D = \alpha_T D_T + \alpha_C D_C \]  (4)

The coefficients \( \alpha_T \) and \( \alpha_C \) are calculated using the following expressions:

\[ \alpha_T = \frac{\Sigma(\varepsilon_{\gamma T})^+}{\Sigma \varepsilon_T^+}; \quad \alpha_C = \frac{\Sigma(\varepsilon_{\gamma C})^+}{\Sigma \varepsilon_C^+} \]  (5)

in which \( \varepsilon_{\gamma T} \) and \( \varepsilon_{\gamma C} \) are determined from the main stresses considering elastic material and \( \varepsilon_T^+ \) represents the total state of elongation, which is equal to:
After the damage variable determination the stress state for the structure (solid) is calculated as follows:

\[ \sigma = (1 - D)D_0 \mathcal{E} \]  

(7)

in which \( D_0 \) is the material elastic properties tensor, \( \mathcal{E} \) is the strain tensor and \( \sigma \) is the stress tensor.

2.2 Steel

2.2.1 Elastoplastic positive isotropic hardening model

One of the characteristics of metallic materials, such as steel, is that this type of material has a plastic behavior after reach the yield stress. Using the plasticity theory it is possible to simulate the appearance and accumulation of residual and plastic deformations. In the hardening phase, there is loss of stiffness, but the material still has strength capacity until its failure limit. Thus, an elastoplastic positive isotropic hardening approach is used to simulate the steel behavior. The criterion used to verify the elastoplastic behaviour is given by:

\[ f = \sigma_s - (\sigma_{sy} + K \zeta) \leq 0 \]  

(8)

in \( \sigma_s \) which indicates the steel reinforcements’ layer stress, \( \sigma_{sy} \) is the steel yielding stress, \( K \) is the hardering plastic modulus and \( \zeta \) is the equivalent plastic strain measurement.

The stress over each reinforcement layer is determined as follows:

\[ f \leq 0 \rightarrow \sigma = E \epsilon \]
\[ f > 0 \rightarrow \sigma = E_t \epsilon \]  

(9)

in which \( E_t \) is the tangent elastoplastic modulus, which is given by: \( E_t = EK/(E + K) \).

2.2.2 Yield stress reduction due to corrosion

An experimental study carried by [8], investigated the residual mechanical capacity of corroded steel bars in concrete. One of the conclusions of the research is that the reduction of yield stress occurs more rapidly than the reduction of cross section steel area.

Based on the results, the authors propose the following expression to predict the yield residual stress of reinforcements after the initiation of the corrosion process:

\[ f_{y,t} = (1.0 - 0.005 Q_{corr}) f_{y,o} \]  

(10)

in which \( f_{y,t} \) and \( f_{y,o} \) are the yield stress of a corroded bar at any time after the despassivation occurs and the yield stress of a non-corroded bar, respectively and, \( Q_{corr} \) is the amount of reinforcements’ corrosion in terms of weight loss, given by the following expression:

\[ Q_{corr} = \frac{0.046 i_{corr}}{a_{initial}} (t - t_r) \]  

(11)
in which the corrosion rate $i_{corr}$ is given in $\mu$A/cm$^2$, the diameter of the non-corroded bar $d_{initial}$ is given in millimeters and the time of corrosion propagation $t$ and to corrosion initiation $t_r$ are given in years.

2.3 Geometric nonlinearity

The updated Lagrangian formulation is used to describe the structure position based on the last equilibrium configuration. Thus, all information required for the next load step is obtained from the last converged load step. The numerical procedure concerns two updates: positions of each structural node and stresses at each integration point along the finite element length. The stress tensor is updated by relating Cauchy’s tensor with the second Piola-Kirchhoff stress tensor. However, for small displacements and strains, Cauchy’s tensor in the current configuration coincides with the second Piola-Kirchhoff tensor of the last equilibrium configuration. Thus, the update process for stress occurs by adding the extra stress of the current step to the last step stress values.

More details about this formulation can be obtained in [9].

2.4 Chlorides transport mechanism

The transport phenomenon associated with the movement of chlorides along structures exposed to aggressive environments is attributed, in most part, to diffusion of chloride ions into concrete pores under a concentration gradient. The coefficient of chloride diffusion, which depends upon the pore structure of concrete, characterizes this flow under a given external concentration of chloride. This parameter is assumed as a characteristic property of hardened concrete.

To simulate the chloride ingress and its transport into concrete pores, Fick's diffusion law [10] has been widely considered as an acceptable model [1, 3, 11]. Fick’s laws for diffusion are applicable for homogeneous, isotropic and inert materials [12].

Considering the mass balance principle, the solution of the differential equation for the especial case where the chloride diffusion coefficient is independent of location, $x$, time, $t$, and chloride concentration, $C$, for a semi-infinite domain with a uniform concentration at the structural surface, is given as follows:

$$C(x, t) = C_0 \text{erfc} \left( \frac{x}{2\sqrt{D_0 t}} \right)$$

in which: $D_0$ is the constant coefficient of diffusion, $C_0$ is the chloride concentration at the structural surface supposed constant in the time and $\text{erfc}$ is the complementary error function.

As one dimensional finite beam element is adopted in this study, it seems to be consistently in representing the chloride ingress in one-dimensional sense. In this work, Eq. (12) is used to evaluate the time in which the corrosion process starts. After the corrosion time initiation, the reinforcements’ areas are penalized using the relations that are presented in next section.

2.5 Reinforcement steel area reduction

In spite of its importance, mathematical approaches for modelling the propagation period of corrosion are not often available in literature, because many researchers consider the initiation period as structural life time. However, the reduction of steel is consistently
represented by models presented in [3, 13, 14]. The models presented on mentioned references were determined empirically, accounting for tropical climate.

In the present study, the steel loss is evaluated using the models presented in [3, 13 14], which assumes that steel is corroded uniformly along its perimeter. For uniform corrosion model, the corroded diameter of reinforcements is determined as follows:

\[
d(t) = \begin{cases} 
d_{\text{initial}} & \text{if } t \leq t_r \\
d_{\text{initial}} - 0.0232 \ i_{\text{corr}} (t - t_r) \geq 0 & \text{if } t > t_r 
\end{cases}
\]  

(13)

in which \(d_{\text{initial}}\) is the reinforcement diameter before depassivation in millimetres, \(t_r\) represents the time of corrosion initiation in years and \(i_{\text{corr}}\) indicates the corrosion rate given in \(\mu A/cm^2\).

The time-dependent corrosion rate is calculated using the expression presented in [14], which was also determined empirically:

\[
i_{\text{corr}}(t) = \left( \frac{37.8(1-w/c)^{-1.64}}{c_{\text{vr}}} \right) 0.85 \cdot t^{-0.29}
\]

(14)

in which \(w/c\) indicates the water/cement ratio and \(c_{\text{vr}}\) represents the concrete cover thickness in cm.

3 STRUCTURAL RELIABILITY

3.1 Direct coupling

In this study, the direct coupling between the finite element mechanical model and the HLRF algorithm to search the design point are used. Required gradients in the HLRF algorithm are calculated numerically through finite differences in the physical space, from the responses of the mechanical model directly.

\[
\frac{\partial F(a, b, ..., z)}{\partial a} = \lim_{\xi_a \to 0} \frac{F(a + \xi_a, b, ..., z) - F(a, b, ..., z)}{\xi_a} \\
\vdots \\
\frac{\partial F(a, b, ..., z)}{\partial z} = \lim_{\xi_z \to 0} \frac{F(a, b, ..., z + \xi_z) - F(a, b, ..., z)}{\xi_z}
\]

(15)

In the present work, \(\xi\) is adopted equal 0.5% of the mean value of the random variable. Thus, this methodology allow to evaluate the limit state equation and their gradients without any approximations.

3.2 Response surface method

A meta model is a model of a model. The creation of a meta model involves the analyze, construction and development of structures, rules, restrictions, models and theories applicable and useful for modelling a particular problem. Among the various existing meta models, the response surface method (RSM) provides good results within structural reliability.

The RSM consist in a simple analytical representation nearby the design point. In this method, the state limit equation is written in a polynomial form. For the construction of the
response surface it is necessary a convenient number of structure simulations. For each simulation held, convenient values are chosen to be assumed by the random variables. This values selection follows a predetermined criterion called experimental plan. An experimental plan is, therefore, a systematic manner to select points where the numerical response will be evaluated.

The polynomial constants are determinate by a regression method. Based on the response at experimental plan points, a polynomial surface is adjusted by using, for example, least square regression:

\[
\min d = \min \sum_{i=1}^{n_p} (Q(x^i) - R^i)^2
\]  

where \(R^i\) is a known response, \(n_p\) is the number of distinct point where response is known and \(Q(x^i)\) is a term of the polynomial response.

3.3 Monte Carlo Simulation Method

Monte Carlo’s method is a numerical simulation procedure widely used in reliability problems. In this method, a sampling of random variables is used to construct a set of values aiming to describe the failure and safe spaces. The sampling is constructed based on the statistical distribution assigned for each random variable considered in the problem. As this method deals with simulation of the limit state function, as bigger be the sampling adopted more accurate will be the spaces’ description and more accurate will be the probability of failure achieved.

The kernel of this method consists on the construction of a sampling for the random variables involved in the problem. The probability of failure is calculated, for Monte Carlo’s simulation, using the following equation:

\[
P_f = \int_D l[x] f_X(x) \, dx = E[l[x]]
\]  

The function \(l[x]\) can be estimated as:

\[
l[x] = \begin{cases} 
1 & \text{if } G(X) < 0 \\
0 & \text{if } G(X) > 0
\end{cases}
\]  

By simulating the limit state function for a convenient number of samplings, the mean value of \(l[x]\) will be an estimator for the probability of failure:

\[
P_f = E[l[x]] = \frac{1}{n_{si}} \sum_{i=1}^{n_{si}} l[x_i] = \frac{n_f}{n_{si}}
\]  

The disadvantage of this method is related to the high number of simulations required in order to compute accurately the probability of failure. However, in the present work the Monte Carlo Simulation is used after the construction of the polynomial response surface. Proceeding in this way, Monte Carlo Simulation will deal only with a simple analytical explicit equation and not with the entire complex mechanical model.

3.4 Definition of limit states and ultimate load

In this paper, analyses are made only for ultimate limit states. The ultimate limit state is configured by the end of strength capacity of a cross section, in which the materials reach
predetermined strain values ($\varepsilon_{c,\text{lim}}$ for concrete and $\varepsilon_{s,\text{lim}}$ for steel) for a given load intensity. The limit state equation for strain in concrete and steel are:

$$
G_{c,ijk} = \varepsilon_{c,\text{lim}} - \varepsilon_{c,ijk}
$$

$$
G_{s,ijkl} = \varepsilon_{s,\text{lim}} - \varepsilon_{s,ijkl}
$$

in which: $i, j, k$ and $l$ are the subscripts, respectively, for the number of finite elements present on the analyses, the number of Gauss points over the length, the number of Gauss point over the cross section height and the number of longitudinal reinforcement layers.

Thus, at the end of each interaction, the strains of materials are verified in every point. The ultimate load is reached when, at any point, the limit state is violated. In case of multiple failure modes, after passing the first failure mode, the second one is characterized when in other section, different of the first one where was found the first failure, the strain of materials reaches any of the strains limit ($\varepsilon_{c,\text{lim}}$ or $\varepsilon_{s,\text{lim}}$). The third, fourth, and so on, ultimate load, are found in an analogous manner.

Mathematically, the ultimate load of a generic failure mode ($n$) is defined by:

$$
P_{\text{ult,mod,n}} = \rho \left\{ \min \left( G_{c,ijk,\text{mod,n}} \leq 0 \right) \cup \left( G_{s,ijkl,\text{mod,n}} \leq 0 \right) \right\}
$$

## 4 APPLICATIONS

As a possible application of the presented methodology, the structure presented in Figure 1 was studied. The configuration of longitudinal reinforcements without any corrosion is shown. The load $P$ is applied vertically on the described nodes, including self-weigh and accidental loads.

The resistant structural load along time was calculated by proposed nonlinear FEM model, with 20 finite elements, using Timoshenko’s theory, physical nonlinearities for concrete, Mazars damage model, and steel, elastoplastic model, and geometrical nonlinearity using updated Lagrangean approach. To obtain internal forces, numerical integration Gauss scheme was adopted, where 6 integration points along element length and 20 integration points its cross section high were used. The following material parameters were adopted: fck equal 30.0 Mpa, longitudinal elasticity modulus of concrete 30672.5 Mpa, coefficient of Poisson 0.2, yield strength of steel 500 Mpa, longitudinal elasticity modulus of steel 196000 Mpa, plastic modulus of steel 19600 Mpa, diameter of stirrups 5.0 mm every 16 cm, concrete cover thickness 40mm and tolerance for convergence based on force norm lesser than $10^{-4}$. The parameters adopted for Mazars damage model were the following: $\varepsilon_{d,0}=6.6102x10^{-5}$, $A_T=0.70$, $B_T=10000$, $A_c=1.5$ and $B_c=2000$. The values used for the evaluation of limit state equation were: $\varepsilon_{c,\text{lim}}=0.0035$ and $\varepsilon_{s,\text{lim}}=0.010$.

To account for chloride ingress and reinforcement corrosion, Fick’s law was used to model chloride penetration and chloride concentration growth along time and the model presented in [3, 13, 14] adopted to represent the steel loss during corrosion. For the analyzes performed, the following parameters were adopted: water/cement ratio 0.5, chloride concentration at structural surface $C_0=1.15$ kg/m³, diffusion coefficient of chlorides $D_0=2.1329x10^{-8}$ cm²/sec, threshold chloride concentration which leads to reinforcements despassivation $C(x, t)=0.90$ kg/m³.
For the reliability analysis two random variables were chosen: fck and the yield strength of steel. The following parameters were adopted for the random variables: fck described by a normal distribution, with mean 30 MPa and standard deviation 3 MPa and the yield strength of steel described by a log normal distribution with mean 500 MPa and standard deviation 50 MPa. The response surface adopted was a complete second grade polynomial and for each corrosion time analyzed 10 million Monte Carlo simulations was realized.

An interesting point of the analysis is about the ultimate loads. The beam’s hyper static level is two and, therefore, it is considered that the beam fails completely when in two distinct points occurs limit state violation. The loads that cause this overcoming of limit state are here called ultimate loads, although a collapse configuration is not presented. The difference founded between the first and second ultimate load value did not result greater than 1% in any increment of time. Therefore, in practical terms, when the first limit state violation load is reached it is very probably that the second one will be reached too.

The beam sequence failure was the same for each increment of propagation corrosion time. Firstly occurred failure of node 11, that is the central support node and then occurred failure of node 17, or node 5 in an equivalent way because the beam is symmetric. The failure, in fact, was expected in these nodes since the central support is the point where the negative bending moment are the greatest all over the beam and the greatest positive bending moment occur in a point near to node 5 or 17.

Figure 2a shows the search process for equilibrium point of node 5 for different instants along the corrosion propagation period. The load P is applied equally to every node of the structure along the search process for equilibrium point. Is possible to note that in the beginning of the loading process all the curves practically coincide. This can be explained by the stiffness provided by concrete to the structure, whereas the reinforcements do not have great influence since they have not yet begun to be required by high stresses. The curves exhibit a behavior ever more distinct as the nodal displacement increases. In conjunction with the advance of the corrosion process, the structure stiffness is compromised. In other words, for a equal load increase the structure begins to have larger displacement and consequently larger strains. Another interesting point is the structural sensibility with respect to the corrosive effects regarding the first ages. In just four years of corrosion propagation the structure shows a reduction of the strength capacity of 20% and in eight years the reduction is 41%. However, after a certain age this reduction is less pronounced. This fact is illustrated.
more clearly in Figure 2b, where is possible to note that from 8 years the reduction of the ultimate load is softened.

![Graph](image)

In the Figure 3 is illustrated the failure probability evolution with the progress of the corrosion process. The failure probability calculated for the first ages, up to 4 years, it was not possible to obtain using the Monte Carlo simulation combined with the RSM. This was not possible, since FORM resulted in a failure probability around $10^{-10}$, i.e., a failure probability value extremely low. To use Monte Carlo simulation, an enormous number of simulations would be needed, making this methodology impractical. From 4 years, the failure probability starts to increase substantially. At age 6, is around $10^{-2}$. This can be explained by the large strength reduction of the structure, since the bars are with severely degraded steel areas, as well as with lower yield strengths. The FORM method was more restrictive about the calculated values for the failure probability over the considered period.

![Graph](image)
5 CONCLUSIONS

There was a rapid decay of strength capacity of the analyzed structure in the first corrosion ages, and in later ages this decay has become less pronounced. Thus, special attention should be given to structures designed in regions with strong environmental aggressiveness.

The proposed model for the analysis of reinforced concrete structures submitted to chloride penetration has the advantage of make natural the search of the failure paths. Failure is not imposed in any structure section, allowing it occur at any point. If the failure occur, it is identified and counted. Thus, in the calculation of the failure probability, is considered the natural path that the structure goes though until it reach the configuration interpreted as “system failure”.

Coupling optimization tools can be an extension of this work, in order to investigate a solution that provides a minimum total cost over the service life of the structure, including initial implementation costs as well as maintenance and repair cost.

Acknowledgement

Sponsorship of this research project by CAPES is greatly appreciated. This research is a part of the activities scheduled by the research project USP/COFECUB 2012.1.672.1.0.

References

APPLICATION OF INTERVENING VARIABLES IN A BILINEAR APPROACH FOR FUZZY STRUCTURAL ANALYSIS

CAMILLO A. PÉREZ* AND MARCOS A. VALDEBENITO*

*Departamento de Obras Civiles, Universidad Técnica Federico Santa María
Av. España 1680, Valparaíso, Chile
e-mail: camilo.perezl@alumnos.usm.cl, marcos.valdebenito@usm.cl

Key words: Fuzzy Analysis, Taylor Series Expansion, Intervening variables.

Abstract. This contribution presents a novel approach for performing fuzzy structural analysis involving linear static problems. The key aspect of the proposed approach is approximating the displacement vector via a first order expansion that involves intervening variables. In addition, linearity of the structural response with respect to load is also exploited in the proposed approach. An example illustrates the advantages of the proposed approach in terms of accuracy.

1 INTRODUCTION

An important and unavoidable issue in modern structural engineering is the correct manipulation and quantification of uncertainties for the aim of achieving robust and reliable designs. This vagueness and imprecise information about a structure is reflected in loads, material properties, cyclical behavior, etc. Nowadays, methods based on fuzzy analysis have gained attention [1] for dealing with such issues. Within fuzzy analysis, each uncertain parameter is characterized through membership functions. Then, through special techniques, the uncertainty on input parameters of a model is propagated to the output, thus allowing to build the membership functions of the response of interest (displacement, stresses, etc.). The main challenge in applying fuzzy structural analysis lies in the impossibility of finding an analytic solution for most cases of practical interest. Hence, one must resort to numerical approaches for uncertainty propagation such as α-level optimization [2], where a series of interval analysis are performed. However, the application of such procedure involves the solution a number of different optimization problems. The solution of these problems demands considerable numerical resources due to the necessity of repeatedly performing structural analyses and possible non-convexity. In order to decrease numerical costs, Taylor series expansion of the response can be constructed [3]. In such way, one single structural analysis is required in order to calculate the solution
for the expansion point and derivatives with respect to fuzzy parameters. Nevertheless, results applying Taylor series expansions are not particularly accurate, especially in problems with high levels of uncertainty.

In this contribution a new approach within the frame of fuzzy structural analysis is presented. The objective is to quantify the variability in form of membership functions of the displacement in linear static structures where the uncertainties are present in both structural properties and loading. This approximation is built from a first order Taylor series expansion [3], incorporating intervening variables to capture the non-linear relationship existing between displacement and structural parameters and direct variables to include the effect of loading. Thus, a high quality approximation for the response of interest is obtained performing a single structural analysis. For the structural parameters, power-type intervening variables are used, involving a real constant that fits the actual non-linear relationship. As an application and in order to highlight the benefits of the presented work, a simple truss structure is analyzed, comparing the obtained results with existing methods and the exact solution, obtained through direct optimization.

2 FORMULATION OF THE PROBLEM

2.1 Fuzzy variables

Consider a linear static structure modeled using the finite element (FE) method. This structure is represented considering $N_p$ uncertain parameters $x_i, i = 1, ..., N_p$ associated with structural properties (e.g. Young’s modulus, cross sectional area, etc.) and $N_l$ uncertain parameters associated with load acting over the structure. Each of the previous variables is modeled through fuzzy variables, i.e.:

\[
\tilde{x}_i = \{(x_i, \mu_{\tilde{x}}(x_i)) : (x_i \in X_i) \land (\mu_{\tilde{x}}(x_i) \in [0, 1])\}, \quad i = 1, ..., N_p
\]

\[
\tilde{y}_j = \{(y_j, \mu_{\tilde{y}}(y_j)) : (y_j \in Y_j) \land (\mu_{\tilde{y}}(y_j) \in [0, 1])\}, \quad j = 1, ..., N_l
\]

where $\tilde{x}_i, i = 1, ..., N_p$ and $\tilde{y}_j, j = 1, ..., N_l$ denote the fuzzy sets modeling the uncertainty in structural properties and load, respectively; $\mu_{\tilde{x}}(x_i), i = 1, ..., N_p$ and $\mu_{\tilde{y}}(y_j), j = 1, ..., N_l$ are the membership functions associated to each uncertain parameter (these functions are assumed convex); $X_i$ and $Y_j$ are the sets that contain $x_i$ and $y_j$, respectively.

In the following, the variables containing uncertainty in structural parameters and loads are grouped in vectors $\mathbf{x}$ and $\mathbf{y}$, respectively.

2.2 Fuzzy structural analysis

Assume the structural system under study is characterized considering a total of $N_d$ degrees-of-freedom. The static equilibrium equation relating structural properties, loads over the system and displacements for the FE model is:

\[
\mathbf{K}(\mathbf{x})\mathbf{u}(\mathbf{x}, \mathbf{y}) = \mathbf{f}(\mathbf{y})
\]
where $K(x)$ is the stiffness matrix of the structural system (dimension $N_d \times N_d$); $u(x, y)$ the displacement vector for the active degrees-of-freedom (dimension $N_d \times 1$) and $f(y)$ the static load vector (dimension $N_d \times 1$). It can be noted that as both stiffness matrix and load vector depend on fuzzy variables, the displacement vector also depends on these variables.

### 2.3 Membership functions and $\alpha$-level optimization

In fuzzy structural analysis the uncertainty in the input parameters is represented as membership functions. In real problems is hard to find an analytic expression for the membership function of the response of interest. Due to this problem, a procedure known as $\alpha$-level optimization is used. This technique consists on performing $N_c$ cuts to the membership functions in order to perform an interval analysis on each level (called $\alpha$ level). Clearly, $0 \leq \alpha \leq 1$. The $\alpha$ level associated to each fuzzy variable is:

\[ x_{i,\alpha_k} = \left\{ x_i \in \tilde{x}_i : \mu_{\tilde{x}_i}(x_i) \geq \alpha_k \right\} , \quad i = 1, \ldots, N_p , \quad \alpha_k \in [0, 1] \]  

\[ y_{j,\alpha_k} = \left\{ y_j \in \tilde{y}_j : \mu_{\tilde{y}_j}(y_j) \geq \alpha_k \right\} , \quad j = 1, \ldots, N_l , \quad \alpha_k \in [0, 1] \]  

where $\alpha_k$, $k = 1, \ldots, N_c$ is the $\alpha$ level under study; $x_{i,\alpha_k}$ and $y_{j,\alpha_k}$ the sets of possible values that $x_i$ and $y_j$ can assume, respectively, for the $\alpha_k$ cut of the associated membership function. Then, the $\alpha$ cut level associated with the $n$-th displacement is:

\[
\mathbf{u}_{n,\alpha_k} = \left\{ u_n : \left( x_i \in x_{i,\alpha_k}, i = 1, \ldots, N_p \right) \land \left( y_j \in y_{j,\alpha_k}, j = 1, \ldots, N_l \right) \land (u_n = u_n(x, y)) \right\} 
\]

\[ n = 1, \ldots, N_d , \quad \alpha_k \in [0, 1] \]  

where $u_n$ is the $n$-th component of the vector $u(x, y)$ that satisfies the equilibrium condition of the equation 3 and $u_{n,\alpha_k}$ is the set of possible values that $u_n$ can assume for the $\alpha_k$ cut in study. Under the assumption that the fuzzy sets $x_{i,\alpha_k}$ and $y_{j,\alpha_k}$ are compact and convex, these can be represented by its lower and upper values (denoted by the superscripts $\cdot^L$ and $\cdot^U$, respectively). Due to the existing relationship between uncertain parameters $(x, y)$ and structural response $u$ (equation 3), the set $u_{n,\alpha_k}$ can be entirely represented by its lower and upper values, denoted by $u_{n,\alpha_k}^L$ and $u_{n,\alpha_k}^U$, respectively. This implies that the determination of the set $u_{n,\alpha_k}$ entails the solution of two optimization problems, i.e. minimize or maximize $u_n$ subject to the restriction that each uncertain parameter lies within a certain interval (for a given $\alpha$-cut) and to satisfy the equilibrium equation. Usually, this is a numerically expensive problem, due to the necessity of repeatedly solving equation 3 (for different combinations of the structural parameters), the existence of local optima, among other problems. Different strategies have been proposed in literature to solve such problem [6]. A possible approach is to
represent the displacement vector by means of an analytic expression constructed using Taylor expansion method; such procedure is studied in the following sections.

3 TAYLOR SERIES EXPANSION

A possible means for approximating the displacement vector \( \mathbf{u}(x, y) \) is applying Taylor expansions \([3]\). These expansions are built from nominal values and partial derivatives evaluated at an expansion point. The first order Taylor series expansion for the displacement vector is:

\[
\mathbf{u}(x, y) \approx \mathbf{u}^L(x, y) = \mathbf{u}(x^0, y^0) + \sum_{i=1}^{N_p} \mathbf{u}_i \left( x_i - x_i^0 \right) + \sum_{j=1}^{N_l} \mathbf{u}_j \left( y_j - y_j^0 \right) \tag{7}
\]

where \((x^0, y^0)\) is the expansion point; \(x_i^0, y_j^0\) are the \(i\)-th \(y\)-th components of the vectors \(x^0\) and \(y^0\), respectively; \(\mathbf{u}(x^0, y^0)\) is the displacement vector evaluated at the expansion point; \(\mathbf{u}_i\) and \(\mathbf{u}_j\) represent the partial derivatives of \(\mathbf{u}\) with respect to \(x_i\) and \(y_j\), respectively and evaluated at the expansion point \((x^0, y^0)\). The vectors \(\mathbf{u}(x^0, y^0)\), \(\mathbf{u}_i\) and \(\mathbf{u}_j\) are calculated through the following expressions:

\[
\mathbf{u}(x^0, y^0) = K^{-1}(x^0) \mathbf{f}(y^0) \tag{8}
\]

\[
\mathbf{u}_i = \frac{\partial \mathbf{u}}{\partial x_i} \bigg|_{x=x^0, y=y^0} = -K^{-1}(x^0) \frac{\partial K}{\partial x_i} \bigg|_{x=x^0} \mathbf{u}(x^0, y^0), \quad i = 1, \ldots, N_p \tag{9}
\]

\[
\mathbf{u}_j = \frac{\partial \mathbf{u}}{\partial y_j} \bigg|_{x=x^0, y=y^0} = K^{-1}(x^0) \frac{\partial \mathbf{f}}{\partial y_j} \bigg|_{y=y^0} \mathbf{u}(x^0, y^0), \quad j = 1, \ldots, N_l \tag{10}
\]

The boundaries for the displacement are obtained minimizing or maximizing equation 7, accordingly. For the \(\alpha\)-level in study, through membership functions it is possible to know the lower and upper limits for each uncertain parameter. Then, depending on the sign of each partial derivative (equations 9 and 10) it is possible to determine the maximum value of the displacement analytically \([3]\). However, the scope of application of such approach is limited due to the incapability of capturing the non-linear relationship existing between displacement and structural parameters, in addition to large errors for high-variability input parameter problems. In order to overcome such issues, a second order Taylor series expansion can be built considering partial derivatives of both first and second order for the displacement vector with respect to structural parameters and loading. Despite of the benefits obtained, this procedure involves the solution of a non-convex optimization problem whose solution is numerically demanding.
4 PROPOSED APPROACH

4.1 Intervening variables

The accuracy of first order Taylor expansion method can be improved in certain cases through the inclusion of intervening variables [5]. These are non linear functions of the uncertain parameters and are defined as follows:

\[ I_i(x_i) = x_i^{m_i}, \quad i = 1, ..., N_p \]  \hspace{1cm} (11)

where \( m_i \) are constants such that \( m_i \in \mathbb{R} \) for \( i = 1, ..., N_p \). The intervening variable of equation 11 has been denominated in the literature of the exponential type. Undoubtedly, the key is the proper choice of the constants \( m_i \) that adapt the relationship between structural displacement and parameters.

4.2 Bilinear implementation

In order to formulate the proposed approach and with the purpose of uncoupling the uncertain loads, assume the load vector is represented as:

\[ f(y) = \sum_{j=0}^{N_l} y_j f_j \]  \hspace{1cm} (12)

where \( f_0 \) is the deterministic part of the load vector (loads without uncertainty); \( y_0 \) is an auxiliary variable such that \( y_0 = 1 \) and \( f_j \) are vectors of dimension \( N_d \times 1 \) that couple the uncertain loads \( y_j \) with the degrees-of-freedom of the structure. In other terms, \( f_j = \frac{\partial f}{\partial y_j} \) for \( j = 1, ..., N_l \). In the following, \( f_j, j = 0, ..., N_l \) are termed as the load participation vectors. In addition, using the former decomposition for the load vector and expressing the structural parameters as intervening variables consider the following definitions:

\[ u_j^0 = K(x^0)^{-1} f_j , \quad j = 0, ..., N_l \]  \hspace{1cm} (13)

\[ u_j^i = -K(x^0)^{-1} \frac{\partial K}{\partial x_i} \bigg|_{x=x^0} \frac{\partial x_i}{\partial I_i} \bigg|_{x_i=x_i^0} u_j^0 , \quad i = 1, ..., N_p, \quad j = 0, ..., N_l \]  \hspace{1cm} (14)

\[ \frac{\partial x_i}{\partial I_i} \bigg|_{x_i=x_i^0} = m_i^{-1} x_i^0 (1 - m_i) , \quad i = 1, ..., N_p \]  \hspace{1cm} (15)

in the former equations, \( u_j^0 \) is the displacement vector for deterministic values of the structural properties and the \( j \)-th load participation vector and \( u_j^i \) is the partial derivative of \( u \) with respect to the intervening variable \( I_i \) considering deterministic values for structural properties and the \( j \)-th load participation vector. Taking into account these definitions, the proposed approach for the displacement vector (denoted as \( u^P \)) is the following.
Camilo A. Pérez and Marcos A. Valdebenito

\[ u(I, y) \approx u^p(I, y) = \sum_{j=0}^{N_l} y_j \left( u_{0,j}^p + \sum_{i=1}^{N_p} u_{i,j}^p (I_i - I_i^0) \right) \]  

(16)

where \( I_i^0 \) is the \( i \)-th intervening variable evaluated at the expansion point. Note the above approximation captures – in some sense – the non linear relationship existing between structural displacement and parameters, involving numerical efforts which are equivalent to those associated to first order Taylor series approach (both require a single factorization of the stiffness matrix). The associated optimization problem (non-convex type) is detailed in the next section.

4.3 Solution for the non-convex optimization problem

With the aim of determining the membership function for the displacement on the \( n \)-th degree of freedom (DOF), the objective is to minimize or maximize the proposed approximation for the displacement. For the degree of freedom of interest, the proposed approximation is:

\[ u^p_n(I, y) = \sum_{j=0}^{N_l} y_j \left( u_{n,0,j}^p + \sum_{i=1}^{N_p} u_{n,i,j}^p (I_i - I_i^0) \right) \]  

(17)

subject to:

\[ I_i \in [I_{i,\alpha_k}^L, I_{i,\alpha_k}^U], \quad i = 1, \ldots, N_p; \quad y_j \in [y_{j,\alpha_k}^L, y_{j,\alpha_k}^U], \quad j = 1, \ldots, N_l \]

where the constants \( u_{n,0,j}^p \) and \( u_{n,i,j}^p \) are obtained evaluating the vectors \( u_{i,j}^p \) and \( u_{i,j}^p \) in the degree of freedom of interest (\( n \)); \( I_i \) and \( y_j \) are the \( i \)-th and \( j \)-th intervening variable and uncertain load, respectively. The minimization or maximization of the above equation corresponds to solving a bilinear optimization problem with disjoint restrictions. This kind of problem has been widely studied in the literature [7] and it has been demonstrated that its global optimum is found to be at the extreme values of each variable. Despite the above, the solution of a bilinear optimization problem is not straightforward. The method used to solve this problem consists in optimizing the relationship between displacement and each variable separately, as performed in first order Taylor series expansion [3]. The algorithm is detailed next.

In the first step, a initial load vector is generated randomly such that each uncertain variable \( y_j, j = 1, \ldots, N_l \) assumes either its lower or upper value (i.e., \( y_{j,\alpha_k} \) assume the value \( y_{j,\alpha_k}^L \) or \( y_{j,\alpha_k}^U \), respectively). In the second step and considering the load vector obtained in the first step, the intervening variables set that minimizes (or maximizes, accordingly) the displacement of interest is identified. Such set involves the lower or upper values of each intervening variable (i.e., \( I_{i,\alpha_k} \) assumes the value \( I_{i,\alpha_k}^L \) or \( I_{i,\alpha_k}^U \)). In the third step, considering the intervening variables vector obtained in the second step, the set of variables \( y_j, j = 1, \ldots, N_l \) that minimize (or maximize, accordingly) the displacement
of interest are identified. Such set involves the lower or upper values of each load variable (i.e. $y_{j,\alpha_k}$ assumes the value $y_{j,\alpha_k}^L$ or $y_{j,\alpha_k}^U$, respectively). In the fourth step, if the uncertain variables $y_j$, $j = 1, ..., N_l$ determined in the first and third step are identical, a local optimum of the optimization problem (minimization or maximization, accordingly) has been found. Otherwise, the algorithm returns to the second step considering the values $y_j$, $j = 1, ..., N_l$, determined in step three. The algorithm is repeated until two cycles produces identical sets of values for the uncertain structural parameters and loads. Despite that the described algorithm does not guarantee convergence to a global optimum, for the studied problems it rapidly converges to the global optimum searched in context with fuzzy analysis.

5 NUMERICAL EXAMPLE, 10-BAR TRUSS STRUCTURE

In order to illustrate the application of the proposed approach, an example involving the estimation of the membership function for the vertical displacement of a 10-bar hyperstatic truss is considered. In this example, the cross sectional area of each bar is modeled as fuzzy variable. The nodal loads applied on this truss are characterized by means of fuzzy variables as well. A schematic representation of the problem and the membership functions that represent the uncertain variables is shown in figure 1. The problem involves a total of 8 DOF’s, and a total of 12 fuzzy variables. The Young’s modulus of each bar is $E = 200$ [GPa]. The objective is to determine the membership function of the vertical displacement in Node 2. Intervening variables of the reciprocal type are used (i.e. $m_i = -1$, $i = 1, ..., 10$). The results obtained are presented in figure 1. In this figure, the membership function is estimated using a first and second order Taylor series expansion (denoted as $T1$ and $T2$, respectively), the proposed approximation (denoted as $P$) and the exact solution (denoted as $E$) obtained through direct optimization. It can be noted that the obtained results are extremely accurate when compared to similar methods.

6 CONCLUSIONS

The implementation of Taylor based expansions allows approximating the displacement of a structural system in a most efficient way. However, as shown in the results of the numerical problem presented in this contribution, the accuracy of such approximation may not be appropriate in context with fuzzy structural analysis. The inclusion of intervening
variables improves the accuracy of Taylor series expansion considerably without incurring into significant additional numerical costs. The key issue for applying intervening variables is the correct choice of this intervening variables, in order to adapt the non-linear relationship existing between structural parameters and displacement.

7 ACKNOWLEDGMENTS

This research is partially supported by CONICYT (National Commission for Scientific and Technological Research) under grant number 11121250. This support is gratefully acknowledged by the authors.

REFERENCES


SURROGATE MODELLING TECHNIQUES IN SOLUTION OF STRUCTURAL RELIABILITY PROBLEMS

HENRIQUE M. KROETZ¹, ANDRÉ T. BECK²

¹Graduate Student, Dept. of Structural Engineering, Univ. of São Paulo, São Carlos, SP, Brazil
Av. Trabalhador são-carlense, nº 400, Centro. CEP 13566-590 São Carlos, SP, Brazil
Henrique.kroetz@usp.br

²Assoc. Prof., Dept. of Structural Engineering, Univ. of São Paulo, São Carlos, SP, Brazil
Av. Trabalhador são-carlense, nº 400, Centro. CEP 13566-590 São Carlos, SP, Brazil
atbeck@sc.usp.br

Key words: Structural Reliability, Uncertainty Quantification, Meta-Models, Artificial Neural Networks, Polynomial Chaos Expansions, Kriging.

Abstract. Solution of structural reliability and uncertainty propagation problems by Monte Carlo simulation can be a demanding task, since complex mechanical models usually have to be solved repeated times. Therefore, surrogate models are often required to reduce the computational burden. This article compares the performance of three surrogate modeling techniques in the solution of structural reliability problems. It addresses artificial neural networks, polynomial chaos and kriging meta-modeling, associated with LHS and Monte-Carlo simulation. A simple procedure for mapping input data for uncertainty quantification problems is also proposed.

1 INTRODUCTION

Structural reliability problems often involve complex mechanical models, usually associated to low failure probabilities. Most computational methods used to solve such problems (e.g. Monte Carlo simulations) require repeated calls of a numerical model of the structural system (e.g. Finite Element, Boundary Element models). Solving reliability problems this way can be a very demanding task, leading to impracticable computational costs. In the past few decades, these difficulties led researchers to search for alternative approaches and opened the so-called meta-modeling field of research.

The first meta-modeling applications in structural reliability involved polynomial response surfaces, and in this context, the original methods were called Response Surface Methods. More advanced techniques have been developed over time, that are able to solve the same problems with more complex, yet greater accuracy surrogate models.

The basic idea is to replace the original detailed numerical model with a substitute, which mimics its behavior, having similar or even identical outputs, for the same input data. In general, a set of support points is selected following some sampling strategy (e.g. randomly, a design of experience, Latin Hypercube sample) and the surrogate model is calibrated using this information. The next step is to solve the problem considering no longer the original model, but the surrogate one.
This paper presents three meta-modeling techniques that have been successfully employed in structural reliability problems, and aims to compare their performance. Artificial Neural Networks (ANN), Polynomial Chaos Expansions (PCE) and Kriging are briefly explained, and then used to solve the same problems, and their accuracy is compared for fixed numbers of support points.

A simple data mapping procedure that has shown to be useful for ANN and Kriging is also discussed.

2 PROBLEM STATEMENT

Let \( X \) be a vector that gathers together the \( m \) random input parameters of the model (e.g. geometrical and material properties, loads) with prescribed density function \( f_X(x) \). The uncertainty implies in the possibility of undesirable structural responses, mathematically described by a limit state function \( g(x) \), such that:

\[
\Omega_f = \{ x | g(x) \leq 0 \} \quad \text{is the failure domain} \tag{1}
\]

\[
\Omega_s = \{ x | g(x) > 0 \} \quad \text{is the survival domain} \tag{2}
\]

The probability of failure of the system is defined by:

\[
P_f = \int_{\Omega_f} f_X(x) \, dx \tag{3}
\]

The integral represented by Eq. (3) usually requires numerical solutions, since its integration domain is implicitly defined, hence the usage of Monte Carlo simulations (MCS). In this technique, the failure probability is interpreted as the mean value of a stochastic experiment where a large number of samples of the random variables are generated [1]. An indicator function \( I(x) \), that assumes value 0 over the survival domain and 1 over the failure domain is considered, and \( n \) samples of \( X \) are generated following \( f_X(x) \). The probability of failure is then estimated as:

\[
P_f = \frac{1}{n} \sum_{i=1}^{n} I(x_i) \tag{4}
\]

Since in MCS convergence is achieved with increasing values of \( n \), it may be useful to evaluate at this point simpler surrogate models instead of the original mechanical models, reducing the computational burden.

3 META-MODELING TECHNIQUES

3.1 Artificial Neural Networks (ANNs)

ANNs are numerical algorithms inspired on the structure of the brain. They have been applied in structural mechanics by [2] and [3], and coupled with MCS in the works of [4], [5] and [6].
A neuron is a logical structure that receives information subjected to characteristic weight functions $w$. In the model adopted in this work, for a neuron $n$ receiving data from $L$ input channels, the neuron weighted input information is subjected to a sum function, resulting in a linear combination of all the received information. A bias term $b$ is added to this result, so that the neuron works with non-zero values even when all the input data are zero. This sum results in the neuron activation value $a$. Equation (5) gives the activation value of a neuron $n$.

$$a_n = \sum_{l=1}^{L} w_{nl} x_l + b_n$$  \hspace{1cm} (5)

The neuron output information is obtained subjecting the activation value to a transfer or activation function $f$, usually an S-shaped function such as hyperbolic, sigmoid, tangent or similar.

Neurons are organized in layers, composing the network. The capability to recognizing patterns, and therefore working as a surrogate model, is defined by the architecture of the network (number of layers and number of neurons per layer), the available training data, and its training algorithm, the later defining weight $w$ for each neuron. Usually, some rule is adopted for initialization of the weights, and an algorithm corrects the weights based on available data about the original model. In the present work, feed-forward networks are used, meaning that the neurons may only send the output information to their next layer. Figure 1 illustrates the operation of a feed forward multi layer neural network.

![Figure 1: Feed forward multi layer neural network.](image)

### 3.2 Polynomial Chaos Expansions (PCE)

PCE has been used since the 90’s in the solution of stochastic finite elements problems, but only in the early 2000’s the expansions began to be seen as meta-models. Broader reviews can be found in the works of [7] and [8].

In this section, the input random variables will be considered independent for a matter of simplicity.
With PCE, the response $Y$ of a structural system is considered as a random variable that belongs to a specific space (e.g. space of random variables with finite variance), and can be represented in a basis of this space, as shown in equation (6):

$$Y = \sum a_\alpha \Psi_\alpha$$  \hspace{1cm} (6)

The $\Psi_\alpha$ are the multivariate polynomial basis functions, that are orthonormal with respect to $f_X(x)$, so that:

$$E[\Psi_\alpha(x),\Psi_\beta(x)] = \delta_{\alpha\beta}$$  \hspace{1cm} (7)

where $\delta_{\alpha\beta} = 1$ if $\alpha = \beta$ and 0 otherwise. The $a_\alpha$ are coefficients to be computed. In the case of independent input random variables, they can be transformed in standard normal random variables, so that $X$ is represented by a standard normal Gaussian vector $\xi = (\xi_1, \xi_2, ..., \xi_m)$. Hermite polynomials are a suitable option for the orthonormal basis functions. For practical computational reasons, the series must be truncated and only the polynomials of degree $p$ or less are adopted. Such basis can be obtained from the tensorization of the truncated univariate expansions of the $m$ input random variables that compose $X$. The number of $P$ basis functions adopted this way is given by:

$$p = \frac{(m + p)!}{m! p!}$$  \hspace{1cm} (8)

A practical algorithm to build the multivariate basis is presented in detail by [9].

The $a_\alpha$ may be obtained non-intrusively by least squares regression. A set of $N$ (at least $P$) support points (realizations of $X$) is selected, and the original model is solved for them, generating the system response vector $Y$. The coefficients are then given by:

$$a = (\Psi^T \Psi)^{-1} \Psi^T Y$$  \hspace{1cm} (9)

where $a$ is a vector that gathers the coefficients $a_\alpha$ and $\Psi$ is the data matrix defined by:

$$\Psi_{ij} = \Psi_{\alpha i}(x_i) / i = 1,...,N; j = 1,...,p-1$$  \hspace{1cm} (10)

With the basis defined and the coefficients determined, the expansion can be used as an approximation for the original model (e.g. a limit state equation).

### 3.3 Kriging

Kriging technique consists in predicting data using interpolation models, which are assembled considering a spatial correlation between the values of the function to be approximated. Kriging has first appeared in the field of geology in the sixties, but only recently, since the works of [10] and [11], the method started to be employed in structural reliability. A historical review of the method is presented by [12].

Kriging models can be seen as the realization of a stochastic field, and thus the function to be surrogated may be written as follows:
\[ g(x) = f(x)^T \beta + z(x) \]  

where the first term defines a deterministic response surface that works as the mean of the process. \( f(x)^T \) is a set of basis functions (e.g. polynomials), and \( \beta \) is a vector that gathers together the coefficients of the surface:

\[ \beta = (F^T R^{-1} F)^{-1} F^T R^{-1} Y \]  

Matrix \( R \) is the correlation matrix, given by equation (15) for each pair of support points, and matrix \( F \) is the regression matrix. The second term of equation (11) is a homogeneous Gaussian random field with zero mean and constant variance \( \sigma_z^2 \):

\[ \sigma_z^2 = \frac{1}{m} (Y - F\beta)^T R^{-1} (Y - F\beta) \]  

It’s autocorrelation function is given by:

\[ Cov[z(x), z(x')] = \sigma_z^2 R(x - x', \theta) \]  

With

\[ R(x - x', \theta) = \prod_{i=1}^{m} e^{-\theta_i d_i^2} \]  

where \( d_i = (x_i - x'_i) \) is the distance between two studied points in the \( i_{th} \) direction. Other autocorrelation functions may also be applied, but the anisotropic Gaussian function is the most common one [13]. Vector \( \theta \) gathers the parameters that define the inverse of the correlation length. Since the correlation matrix \( R \), hence \( \beta \) and \( \sigma_z^2 \), depend on \( \theta \), the later vector has to be first calculated. This is done by minimizing the function \( \mathcal{L}(\theta) \), using the method of maximum likelihood:

\[ \mathcal{L}(\theta) = |R(\theta)|^{-\frac{1}{2}} \sigma_z^2(\theta) \]  

Once \( \theta \) is defined, all the kriging parameters can be obtained, and a limit state function can be finally predicted by equation (11), with:

\[ z(x) = r(x)^T R^{-1} (Y - F\beta) \]  

And

\[ r_i(x')^T = [R(x - x'_i, \theta)] \]

4 DATA MAPPING

The performance of many surrogate modeling techniques is severely affected by the format assumed by their input parameters. In structural reliability, for instance, often input random variables have very different orders of magnitude (e.g. Young Modulus and cross-section area of a beam when both are expressed in the International System of Units). Therefore, data must receive some treatment, to ensure that the approximation is satisfactorily accurate. With PCE, this is intrinsically made when vector \( X \) is transformed into \( \xi \), since standard normal variables assume only values reasonably close to the unity. For polynomial response surfaces and Kriging, on the other hand, matrix inversions must be performed, since crude data may result in ill-posed matrixes. For ANNs, the input values are subjected to
activation function $f$, but its S-shaped nature implies on its derivative being practically constant when the input range is too large, and this disrupts the efficiency of many training algorithms.

Since in structural reliability problems the input variables are random ones, their moments information can be used in the mapping. Even though an experienced user may input the information in compatible units, some general procedure may be useful, especially when a computer program is being developed. In the present work, the input random variables were divided by their mean before feeding the models. In the few cases where the mean was zero, input random variables were divided by their standard deviation. This has shown to be enough for overcoming the aforementioned problems.

5 Application Examples

5.1 Some details on the employed Meta-Models

For every application, a Multi-Layer Perceptron ANN is adopted with only one hidden layer (as shown to be enough by [14]). The first layer has one neuron for each input random variable and the last layer has one neuron for each response of the model. The number of neurons of the hidden layer is twice the number of input parameters plus one, as stated by the Kolmogorov–Nielsen theorem. A tangent-sigmoid transfer function was adopted for the neurons on the hidden layer, and linear functions on the other ones. When the number of training points is stated on the examples, it already considers the 20% of the information that was used as validation data. The Levenberg–Marquardt method was used to train the networks, with 2000 iterations. More details about this training method can be found in [15].

Two full expansion PCEs are used in the analytical examples, with the max (multi-index) polynomial degree of 4 and 7. In section 4.3, a full PCE expansion of degree 3 is considered, same order used by [16] in the same example. The PCE coefficients were obtained with the regression approach discussed in section 2.2. In Kriging models, the correlation is defined by an isotropic Gaussian correlation function, and the sample support points are centered on the random variable’s means. For the response surface, up to second order polynomials with no cross-terms were considered as basis functions.

In all cases, the support/training points were defined by a random Latin Hypercube Sampling technique. Once the surrogate models are established, crude Monte Carlo simulation with $10^6$ samples is used on the original model as a reference, and then for each analytical technique. Only $10^4$ samples were used in the numerical example, since the accuracy of the meta-models, and not the exact solution, is studied.

5.2 Analytical Limit State Examples

Four limit state equations that can be analytically represented are considered in this section. They are solved for 2 different numbers of support points: the minimum required for the full $7^{th}$ degree PCE, as given by equation 8 (i.e. the minimum required for all the techniques to work) and twice this number. Eventually, a few support points were added so that data matrix became well conditioned. The random variables used in this section are listed in Table 1. Table 2 shows the limit state equations addressed in this section.
Table 1: Random Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1, X_2, X_3$</td>
<td>Standard Normal</td>
</tr>
<tr>
<td>$X_4, X_5, X_6$</td>
<td>Uniform$\sim[-\pi, \pi]$</td>
</tr>
</tbody>
</table>

Table 2: Studied Limit State Equations.

<table>
<thead>
<tr>
<th>Eq. Type</th>
<th>$g(x)$</th>
<th>Eq. #</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>$2X_1 - X_2 + 5$</td>
<td>(19)</td>
</tr>
<tr>
<td>Polynomial</td>
<td>$4X_1 + 2X_2^2 - X_1X_3^3 + 10$</td>
<td>(20)</td>
</tr>
<tr>
<td>Exponential</td>
<td>$e^{(0.2X_1 + 1.4)} - X_2$</td>
<td>(21)</td>
</tr>
<tr>
<td>Ishigami</td>
<td>$\sin X_4 + 7 \sin^2 X_5 + 0.1X_6^4 \sin X_4 + 8$</td>
<td>(22)</td>
</tr>
</tbody>
</table>

The reliability index $\beta$ obtained with each technique is shown in Table 3, for the minimum required number of training points, and in Table 4 for the models trained with twice this number. A positive constant was added to equation (22) to maintain its complexity but make it more suitable for the limit state equation behavior of a hypothetical structural reliability problem.

The used meta-models appear to satisfactorily represent all the studied analytical equations, with improved accuracy as the number of support points grow. Some difficulties were found when dealing with the ANN, for slightly changing the number of support points implied in a considerable gain or loss of accuracy on the reliability index. Also, for PCE, the data matrix often became ill-posed, the selection of a different number of support points being required for an accurate solution. Still, in both cases the errors appear to be local effects, the accuracy still increasing with the number of support points, taken away the punctual errors.

Table 3: $\beta$ for minimum number of support points.

<table>
<thead>
<tr>
<th>Eq.</th>
<th>Ref.</th>
<th>NNA</th>
<th>$PCE_4$</th>
<th>$PCE_7$</th>
<th>Kriging</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>2,23</td>
<td>2,23</td>
<td>2,23</td>
<td>2,23</td>
<td>2,23</td>
</tr>
<tr>
<td>20</td>
<td>2,11</td>
<td>2,07</td>
<td>2,11</td>
<td>2,11</td>
<td>2,11</td>
</tr>
<tr>
<td>21</td>
<td>3,36</td>
<td>3,38</td>
<td>3,36</td>
<td>3,36</td>
<td>3,36</td>
</tr>
<tr>
<td>22</td>
<td>1,00</td>
<td>1,00</td>
<td>0,88</td>
<td>0,91</td>
<td>1,00</td>
</tr>
</tbody>
</table>

Table 4: $\beta$ for more support points.

<table>
<thead>
<tr>
<th>Eq.</th>
<th>Ref.</th>
<th>NNA</th>
<th>$PCE_4$</th>
<th>$PCE_7$</th>
<th>Kriging</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>2,23</td>
<td>2,23</td>
<td>2,23</td>
<td>2,23</td>
<td>2,23</td>
</tr>
<tr>
<td>20</td>
<td>2,11</td>
<td>2,11</td>
<td>2,11</td>
<td>2,11</td>
<td>2,11</td>
</tr>
<tr>
<td>21</td>
<td>3,36</td>
<td>3,36</td>
<td>3,36</td>
<td>3,36</td>
<td>3,36</td>
</tr>
<tr>
<td>22</td>
<td>1,00</td>
<td>1,00</td>
<td>0,99</td>
<td>0,99</td>
<td>1,00</td>
</tr>
</tbody>
</table>
5.2 Finite Elements Truss

A truss structure first studied by [17] is considered, as represented by figure 2. It is composed by 23 bars and 13 nodes, and it is subjected to 6 vertical random loads on the upper nodes. The input random parameters are described in Table 5.

Figure 2: Studied Truss Structure.

The limit state equation is defined implicitly as the result of a finite element model, taking into consideration the vertical displacement of the mid-span node, as represented in Figure 2. The serviceability of the truss relates to a threshold of 10cm defined as limit displacement, thus the limit state equation reads:

\[ g(x) = 10 - |V_1(x)| \]  \hspace{1cm} (23)

The reliability indexes obtained with each technique for various numbers of support points are plotted in Figure 3.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Distribution</th>
<th>Mean</th>
<th>St. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>( E_1, E_2 ) (Pa)</td>
<td>Lognormal</td>
<td>( 2.1 \times 10^{11} )</td>
<td>( 2.1 \times 10^{10} )</td>
</tr>
<tr>
<td>( A_1 ) (m²)</td>
<td>Lognormal</td>
<td>( 2 \times 10^{-3} )</td>
<td>( 2 \times 10^{-4} )</td>
</tr>
<tr>
<td>( A_2 ) (m²)</td>
<td>Lognormal</td>
<td>( 1 \times 10^{-3} )</td>
<td>( 1 \times 10^{-4} )</td>
</tr>
<tr>
<td>( P_1 - P_6 ) (N)</td>
<td>Gumbel</td>
<td>( 5 \times 10^{4} )</td>
<td>( 7.5 \times 10^{3} )</td>
</tr>
</tbody>
</table>
Figure 3: β for different number of support points

All the techniques converge to the reference value of $\beta = 1.71$ as the number of support points increase. The full PCE expansion shows good convergence, but it requires use of at least 286 support points, while ANN and Kriging give reasonably accurate results with as few as 50 support points. The fastest convergence was presented by the ANNs, even though the training process may be time demanding in comparison to the definition of the kriging meta-model.

6 CONCLUSIONS

Three meta-modeling techniques were employed in the solution of five different problems. Their accuracy has been compared, and in all cases found to be satisfactory. Their convergence with increasing number of support points has been verified in all the studied examples. A suggestion for the often required data mapping was discussed in detail. Some difficulties which may be found when dealing with each technique were pointed out, and a few more are briefly addressed here:

- In the regression performed by PCE and Kriging, the data matrix may, in some cases, not be invertible. Regularization of the matrix by adding small quantities to its main diagonal, using a different number of support points and replacing the inverse matrix by the corresponding pseudo-inverse are ways of overcoming this problem.

- For mechanical models with very large number of random input parameters, use of PCE may become impracticable, since the minimum number of times the original model must be called is too large. The sparse PCE approach presented in [18] mitigates this difficulty.

- The ANNs showed great accuracy for all studied examples, but was by far the most time consuming technique. Also, the division between training and validation data may change this models accuracy, even for the same total number of support points.

Further work is in progress towards addressing more challenging structural reliability problems and employing the studied techniques in solution of risk optimization problems.
REFERENCES


SFEM USING A VOLUMETRIC-DEViatorIC SPLIT OF THE ELASTICITY TENSOR

I. CAYLAK, R. MAHNKEN AND A. DRIDGER

University of Paderborn, Chair of Engineering Mechanics (LTM)
Warburger Strasse 100, 33098 Paderborn
caylak@ltm.upb.de

Key words: Computational Mechanics, FEM, SFEM, Polynomial Chaos

Abstract. This paper investigates the uncertainty of a linear elastic body by modeling the elastic random material parameters as stochastic fields. For its stochastic discretization a polynomial chaos (PC) is used to expand the coefficients in deterministic and stochastic parts. The key idea of our contribution is the split of the linear elasticity tensor into a volumetric and deviatoric part. Then, from experimental data for an adhesive material the distribution of the random variables, i.e. Young’s modulus \( E(\omega) \) and the shear modulus \( G(\omega) \), are known. Consequently, \( E(\omega) \) and \( G(\omega) \) are expanded with PC, and a Galerkin projection \[1\] can be applied to reduce the stochastic partial differential equations (SPDEs) into a system of deterministic partial differential equations (PDEs). Dependent random material parameters with two resulting PC coefficients are considered. In the numerical example a plate with a ring hole under tension is investigated. The probability density of the displacement obtained by the PC based stochastic finite element is compared to the probability density obtained by the Monte Carlo (MC) simulation. Both methods, PC and MC reveal similar results with small deviations.

1 INTRODUCTION

In many engineering applications the materials are heterogeneous. Typical examples are adhesives, polycrystallines and composites. This heterogeneity leads often to uncertainty in the material parameters and to uncertainty in the mechanical response. Therefore, macroscopically heterogeneous materials should be modelled by a stochastic approach instead a deterministic approach. Mathematically, the mechanical system with random material parameters can be described by SPDEs with stochastic fields, which are solved by the stochastic finite element method (SFEM). Mostly used methods in SFEMs are Monte Carlo (MC), Galerkin and Polynomial Chaos Expansion (PCE) \[3],[5\].

The aim of this work is to consider the uncertainty of a linear elastic body. The uncertainty is considered by random material parameters, which are modeled as stochastic.
fields. Therefore, unknown stochastic fields are expanded with the Polynomial Chaos method and a Galerkin approach is used to solve the associated unknown coefficients. The key idea of our contribution is the split of the linear elasticity tensor into a volumetric and deviatoric part. Then, from experimental data for an adhesive material the distribution of the random variables, i.e. Young’s modulus $E(\omega)$ and the shear modulus $G(\omega)$, are known. Consequently, $E(\omega)$ and $G(\omega)$ are expanded with the PCE, and a Galerkin projection [1] can be applied to reduce the SPDE into a system of deterministic PDEs. Finally, the finite element matrix equation consists of two contributions, namely the deterministic and the variation part.

As a numerical example we consider the static problem for uniaxial tension of the rectangular plate with a ring hole. This problem is investigated under plane strain conditions. Results of the deterministic solution and the influence of the distribution of the material parameters $E(\omega)$ and $G(\omega)$ on the solution are presented.

2 Governing equations
2.1 Strong formulation

We consider a linear elastic body problem in a state of small strains that occupies the domain $\mathcal{B} \subset \mathbb{R}^3$. Furthermore, the probability space is denoted by $(\Omega, \Sigma, P)$, where $\Omega$ is the set of elementary events, $\Sigma$ is the $\sigma$-algebra and $P$ is the probability measure. Let $\omega$ be an element of $\Omega$. The material behavior is described by two fluctuating material parameters, namely Young’s modulus $E(\omega)$ and the shear modulus $G(\omega)$, respectively. Using standard tensor notation, the field equations read

$$\nabla \cdot \sigma(\omega) + b = 0, \quad (1)$$
$$\sigma(\omega) = C(\omega) : \varepsilon(\omega), \quad (2)$$
$$\varepsilon(\omega) = \frac{1}{2}(\nabla u(\omega) + (\nabla u(\omega))^T). \quad (3)$$

Here in Eq.(1) $b$ is the body force per unit volume and $\sigma(\omega)$ denotes the Cauchy stress tensor, where the constitutive relation is described in Eq.(2). We consider a deviatoric-volumetric split of the elasticity tensor as

$$C(\omega) = (C_{\text{dev}}(\omega) + C_{\text{vol}}(\omega)) = 2G(\omega)I_{\text{dev}} + \kappa(\omega) 1 \otimes 1. \quad (4)$$

In Eq.(4) the parameter $\kappa(\omega)$ could be calculated using the parameters $E(\omega)$ and $G(\omega)$. For the calculation of $\kappa(\omega)$ we have to multiply and divide stochastic variables which is obviously not trivial and will be explained in a later section. In Eq.(3) $\varepsilon(\omega)$ is the strain tensor obtained from the symmetry gradient of the displacement field $u(\omega)$.

2.2 Weak formulation

For construction of a weak formulation, we take the scalar product of Eq.(1) with a test function $\delta u(\omega)$, integrate over the volume $\mathcal{B}$ apply the Gauss theorem and impose
traction boundary condition. This procedure results into the weak formulation
\[
a(\delta u(\omega), u(\omega)) - G_{ext} = 0, \quad \text{(5)}
\]
where
\[
a(\delta u(\omega), u(\omega)) = \int_{\Omega} \int_{\partial B} \delta \epsilon(\omega) : (C_{\text{dev}}(\omega) + C_{\text{vol}}(\omega)) : \epsilon(\omega) dV dP,
\]
\[
G_{ext} = \int_{\Omega} \int_{\partial B} \delta u(\omega) \cdot b dV dP + \int_{\Omega} \delta u(\omega) \int_{\partial B} \bar{t} dAdP. \quad \text{(7)}
\]
The continuous bilinear form in Eq.(5) could be solved by the stochastic finite element method, where the Galerkin method is used to convert the continuous problem to a discrete problem. Here, we have to consider a spatial discretization for the classical part and a stochastic discretization for the random fields \(\delta u(\omega), \delta \epsilon(\omega), C_{\text{dev}}(\omega)\) and \(C_{\text{vol}}(\omega)\) in Eq.(5)-(7).

3 Spatial discretization and finite element matrix formulation

We consider a spatial discretization of the domain \(B \approx B^h = \bigcup_{e=1}^{n_e} B_e\) into \(n_e\) elements, see more details in [7]. Discretization of the weak formulation Eq.(5)-Eq.(7) and using interpolation functions renders after some rearrangements the finite element matrix formulation

\[
K(\omega)u(\omega) = F_{ext}, \quad \text{(8)}
\]
\[
K(\omega) = K_{\text{dev}}(\omega) + K_{\text{vol}}(\omega), \quad \text{(9)}
\]
\[
K_{\text{dev}}(\omega) = A_{e=1}^{n_e} \int_{B_e} B^T C_{\text{dev}}(\omega) B dV, \quad \text{(10)}
\]
\[
K_{\text{vol}}(\omega) = A_{e=1}^{n_e} \int_{B_e} B^T C_{\text{vol}}(\omega) B dV, \quad \text{(11)}
\]
\[
F_{ext} = A_{e=1}^{n_e} \int_{\partial B_e} N^T \bar{t} dA. \quad \text{(12)}
\]
Here in Eq.(8) \(K(\omega)\) is the system stiffness matrix which is calculated with Eq.(9) as a split into a deviatoric and volumetric part. Additonally, we introduce the assembly operator \(A\).

4 Polynomial Chaos and discretization of random fields

4.1 PC representation

In this paper the random variables are expanded with the PCE. The PCE involve a basis of known random functions with deterministic coefficients, which is of course an advantage compared e.g. with the Karhunen-Loeve Expansion (KLE). Then, the PCE of an arbitrary random variable \(X(\omega)\) can be formulated as

\[
X(\omega) = \sum_{i=1}^{\infty} \hat{X}_i \psi_i(\xi(\omega)), \quad \text{(13)}
\]
where $\hat{X}_i$ are deterministic PC coefficients. The PC basis functions $\psi_i(\xi(\omega))$ are described by Hermite polynomials with standard normally distributed random variables $\omega$. The first five Hermite polynomials have the form
\[
\begin{align*}
\psi_1 &= 1; \\
\psi_2 &= \xi; \\
\psi_3 &= \xi^2 - 1; \\
\psi_4 &= \xi^3 - 3\xi; \\
\psi_5 &= \xi^4 - 6\xi + 3.
\end{align*}
\]
(14)

Due to the orthogonality property of Hermite polynomials it follows, see [4]
\[
\begin{align*}
\mathbb{E}(\psi_1) &= 1, \\
\mathbb{E}(\psi_i \psi_j) &= \delta_{ij}, \\
\mathbb{E}(\psi_i) &= 0 \quad \forall i \neq 1,
\end{align*}
\]
(15)
where $\mathbb{E}(\psi_i)$ denotes the expected value of $\psi_i$ and is defined for an arbitrary random variable as, see [1]
\[
\mathbb{E}(X) = \int_{\Omega} X d\mathbb{P} = \int_{\Omega} X(\omega)d\mathbb{P}(\omega).
\]
(16)

An exact PC representation of an arbitrary random variable is given in Eq.(13). The discretization of $X(\omega)$ can be realized by limiting the number of PC coefficients and PC basis
\[
X(\omega) \approx \sum_{i=1}^{P} \hat{X}_i \psi_i(\xi(\omega)).
\]
(17)

Finally, elementary arithmetic operations on PC variables is presented. The basic operations on PC variables in this work are addition, multiplication and division of two random variables $X(\omega)$ and $Y(\omega)$

\textbf{Addition:} \quad Z(\omega) = X(\omega) + Y(\omega) \quad \Rightarrow \hat{Z}_i = \hat{X}_i + \hat{Y}_i
\]
(18)

\textbf{Multiplication:} \quad Z(\omega) = X(\omega)Y(\omega) \quad \Rightarrow \hat{Z}_k = \sum_{i=1}^{P} \sum_{j=1}^{P} \frac{\mathbb{E}(\psi_i \psi_j \psi_k)}{\mathbb{E}(\psi^2_i)} \hat{X}_i \hat{Y}_j
\]
(19)

\textbf{Division:} \quad Z(\omega) = \frac{X(\omega)}{Y(\omega)} \quad \Rightarrow \hat{Z}_k = \left( \sum_{i=1}^{P} \sum_{j=1}^{P} \frac{\mathbb{E}(\psi_k \psi_j \psi_i)}{\mathbb{E}(\psi^2_i)} \hat{Y}_j \right)^{-1} \hat{X}_k
\]
(20)

In Eq.(18)-(20) the results for arithmetic operations adding, multiplying and dividing on PC variables are given. For specific explanations we refer to [2].

\section{Calculation of PC coefficients based on experimental data}

The stochastic collacation method [2] is one possibility to calculate the deterministic PC coefficients $\hat{X}_i$ of a random variable $X(\omega)$. From experimental investigations the distribution of the random variables, i.e. Young’s modulus $E(\omega)$ and the shear modulus
$G(\omega)$, are known for an adhesive material. The general framework for a statistical analysis of material parameters is described in [6]. Based on ten experimental stress-strain curves, in combination with a non-linear regression function and an autoregressive moving average process, artificial data sets are generated. For all artificial data sets a parameter identification is performed and a distribution of each parameter is obtained. In Figure 1 a) and b) the probability density function (PDF) of the material parameters $E(\omega)$ and $G(\omega)$ from experiments are given. In the first step the deterministic PC coefficients of

$$E(\omega) = \hat{E}_1 + \hat{E}_2 \xi \quad \text{and} \quad G(\omega) = \hat{G}_1 + \hat{G}_2 \xi,$$

(21)

with the deterministic PC coefficients $\hat{E}_1 = 1425.58 \text{MPa}$, $\hat{E}_2 = 136.67 \text{MPa}$, $\hat{G}_1 = 543.06 \text{MPa}$ and $\hat{G}_2 = 31.61 \text{MPa}$. Then, using elementary arithmetic operations on PC variables given in Eq.(18)-(20) in the second step with $E(\omega)$ and $G(\omega)$ from Eq.(21) the compression modulus $\kappa(\omega)$ is calculated

$$\kappa(\omega) = \frac{G(\omega)E(\omega)}{3(3G(\omega) - E(\omega))} = \hat{\kappa}_1 + \hat{\kappa}_2 \xi.$$

(22)

The deterministic PC coefficients of $\kappa(\omega)$ are $\hat{\kappa}_1 = 1380.99 \text{MPa}$, $\hat{\kappa}_2 = 500.94 \text{MPa}$, and the results are illustrated in Figure 1 c). Then, inserting Eq.(21) and Eq.(22) into Eq.(9)
the deterministic PC coefficients of system stiffness matrix $K(\omega)$ could be obtained as

$$K(\omega) = \hat{K}_1 + \hat{K}_2 \xi,$$

(23)

$$\hat{K}_1 = \sum_{e=1}^{ne} \int_{\mathcal{E}_e} B^T (\hat{C}_{dev1} + \hat{C}_{vol1}) B dV,$$

(24)

$$\hat{K}_2 = \sum_{e=1}^{ne} \int_{\mathcal{E}_e} B^T (\hat{C}_{dev2} + \hat{C}_{vol2}) B dV,$$

(25)

$$\hat{C}_{dev1} + \hat{C}_{vol1} = 2 \hat{G}_1 \mathbf{I}_{dev} + \hat{\kappa}_1 \mathbf{1}_1^T,$$

(26)

$$\hat{C}_{dev2} + \hat{C}_{vol2} = 2 \hat{G}_2 \mathbf{I}_{dev} + \hat{\kappa}_2 \mathbf{1}_1^T.$$

(27)

5 Stochastic finite element matrix formulation

In this section the SFEM will be derived based on a Galerkin method. To this end the stochastic discretization of (17) will be used in order to represent the PC formulation of the stochastic variables $K(\omega)$ and $u(\omega)$

$$K(\omega) = \hat{K}_1 + \hat{K}_2 \xi \quad \text{and} \quad u(\omega) = \hat{u}_1 + \hat{u}_2 \xi,$$

(28)

Afterwards, they will be inserted into the finite element matrix formulation in Eq.(8)

$$\sum_{i=1}^{P} \hat{K}_i \psi_i \sum_{j=1}^{P} \hat{u}_j \psi_j = F_{ext}.$$

(29)

Then, the stochastic Galerkin method will be applied. This requires the orthogonality between Eq.(23) and $\psi_k$ regarding the expected value which renders the stochastic finite element equation

$$\begin{bmatrix} \hat{K}_{11} & \hat{K}_{12} \\ \hat{K}_{21} & \hat{K}_{22} \end{bmatrix} \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \end{bmatrix} = \begin{bmatrix} F_{ext} \\ 0 \end{bmatrix},$$

(30)

where $\hat{K}_{jk}$ in Eq.(30) is $\sum_{i=1}^{P} \mathbb{E}(\psi_i \psi_j \psi_k) \hat{K}_i$. As described in Section 4.2 we use normally distributed experimental data. Therefore, the number of PC coefficients is limited to $P = 2$. Then, Eq.(30) could be expressed as

$$\begin{bmatrix} \hat{K}_1 & \hat{K}_2 \\ \hat{K}_2 & \hat{K}_1 \end{bmatrix} \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \end{bmatrix} = \begin{bmatrix} F_{ext} \\ 0 \end{bmatrix},$$

(31)

where $\hat{K}_1$ and $\hat{K}_2$ are the PC coefficients of Eq.(24) - Eq.(25).
6 Numerical example

In the numerical example a plate with a ring hole under tension will be investigated. The geometry and the finite element discretization of this two-dimensional problem are shown in Figure 2. Due to obvious symmetry conditions only a quarter of the specimen is discretized with 1328 triangular elements with the resulting mesh of Figure 2b). The distributed load at the top is \( t = 200 \text{ MPa} \). The material parameters are \( \hat{G}_1 = 543.06 \text{ MPa} \) and \( \hat{G}_2 = 31.61 \text{ MPa} \) for the shear modulus and \( \hat{\kappa}_1 = 1380.99 \text{ MPa} \), \( \hat{\kappa}_2 = 500.94 \text{ MPa} \) for the compression modulus. Figure 3 illustrates the displacement fields of plate in \( x \)- and \( y \)-direction, where Figure 3a) and 3c) represent the mean values \( \hat{u}_{1x} \) and \( \hat{u}_{1y} \). The corresponding standard deviations \( \hat{u}_{2x} \) and \( \hat{u}_{2y} \) are given in Figure 3b) and 3d). Regarding the percentage of the standard deviation in \( x \)- and \( y \)-direction we could observe that a pronounced standard deviation in \( y \)-direction exists, where \((\hat{u}_{2x})\% 9\% \) and \((\hat{u}_{2y})\% 13\% \). Additionally, in Figure 4 the probability density of the displacement at node A in \( x \)-direction obtained by PC based stochastic finite element is compared to the probability density obtained by the Monte Carlo (MC) simulation. Both methods, PC and MC reveal similar results with small deviations. The errors in the PC coefficients could be quantified by

\[
E_{PC,MC} = \sum_{i=1}^{n_{exp}} (\mathbf{u}_{PC} - \mathbf{u}_{MC})^2. \tag{32}
\]

\( \mathbf{u}_{MC} \) is the solution vector of \( n_{exp} \) computations with the material parameters from experiments and \( \mathbf{u}_{PC} \) is the deterministic solution multiplied \( n_{exp} \) times with randomly generated standard normally values. Resulting values for errors at node A in \( x \)-direction and node B in \( y \)-direction are \( E_{PC,MC} = 0.142 \) and \( E_{PC,MC} = 0.014 \), respectively. The deviation occurs due to the fact that the probability density of the material parameters is not exactly normally distributed which can be seen in Figure 1. In order to reduce the error between PC and MC, the number of PC coefficients must be increased. This behavior will be published in a separate paper, which is currently in preparation. Finally, we compare the computational cost of MC and PC, where the PC simulation only needs 5.4 seconds and the MC simulation needs 2,558.8 seconds.

7 Summary and Conclusions

This paper presents a PC based stochastic finite element method in order to describe the linear elastic response of an adhesive with two random material parameters. The distribution of the Young's modulus \( E(\omega) \) and the shear modulus \( G(\omega) \) are given from experiments. Then, using the collocation method the PC coefficients for both distributions are determined. Furthermore, the elasticity tensor is divided into a deviatoric and volumetric part. The deviatoric part is described by \( G(\omega) \), where the volumetric part needs \( \kappa(\omega) \). Using elementary arithmetic operations on PC variables the PC coefficients of \( \kappa(\omega) \) are calculated. Applying a spatial and stochastic discretization of the weak formulation leads to a stochastic finite element formulation with two PC coefficients for the solution.
Figure 2: Plate with a ring hole: a) Geometry and b) finite element mesh of triangular elements

Figure 3: Plate with a ring hole: a) $\hat{u}_{1x}$ and b) $\hat{u}_{2x}$ in $x$-direction, c) $\hat{u}_{1y}$ and d) $\hat{u}_{2y}$ in $y$-direction
In the numerical example a plate with a ring hole under tension is investigated. The probability density of the displacement obtained by PC based stochastic finite element is compared to the probability density obtained by the MC simulation. Both methods, PC and MC reveal similar results with small deviations. Furthermore, the main advantage of the PC method due to less computational cost is shown.

REFERENCES


DYNAMIC BEHAVIOR OF TIMBER BEAMS WITH UNCERTAIN PROPERTIES

DIEGO A. GARCÍA¹, MARTA B. ROSALES² AND RUBENS SAMPAIO³

¹,²Department of Engineering, Universidad Nacional del Sur.
Alem 1253, 8000 Bahía Blanca, Argentina.
CONICET, Argentina.
garciadiego@fio.unam.edu.ar mrosales@criba.edu.ar

³PUC-Rio, Department of Mechanical Engineering, Pontificia Universidade Católica.
Rua Marquês de São Vicente 225, 22453-900 Rio de Janeiro, Brazil.
rampaio@puc-rio.br

Key words: Timber Beams, Uncertain Properties, Modal Analysis, Forced Vibrations.

Abstract. The dynamic behaviour of sawn timber beams of Argentinian Eucalyptus grandis using a stochastic model is herein presented. In first place, the free vibrations case is studied. After solved, the natural frequencies and modes within the statistical frame. Finally, the damped forced vibration of the structural element is addressed. For the first problem, the governing differential equations are stated starting from the Euler-Bernoulli (E-B) beam theory. The lengthwise variability of the Modulus of Elasticity (MOE) and of the second moment of the cross sectional area are simulated to account for the presence of knots. The variability of the mass density among structural elements also is considered. For the second problem, the governing differential equations are stated starting from the Timoshenko beam theory. The damping is introduced through the elastic and shear complex moduli. This material damping model provides us of an adequate description of the energy dissipation under the action of harmonic loads. Uncertainties in this problem are introduced in the elastic and shear moduli and their corresponding bending and shear loss factor, respectively. The governing equations are discretized and numerically approximated through the Finite Element Method (FEM). Then, the statistics of the solutions are obtained. The Probability Density Functions (PDFs) of the natural frequencies and the FRF are numerically obtained through the Stochastic Finite Element Method (SFEM) and Monte Carlo Simulations (MCS). Numerical results of the first natural frequency are validated with experimental values. Also, the mode shapes statistics are analyzed. The problems herein studied contribute to attain a more realistic description of the dynamic behavior of structures made out of sawn timber.
1 INTRODUCTION

Due to its natural origin, structural timber is characterized by considerable variability of its mechanics properties. Timber knots are the main source of the lengthwise variability of bending strength and stiffness in sawn beams. It is apparent that a stochastic approach becomes necessary in order to attain a more realistic structural model.

A simple method for visually strength grading sawn timber of Argentinian *Eucalyptus grandis* has been developed by Piter [1] and they conclude that the presence of pith and knots are considered the most important visual characteristic for strength grading this material by the Argentinian standard IRAM 9662-2 [2].

Czmoch [3] studied the bending strength in sections with knots and determined the load carrying capacity of timber beams with knots through a Poisson process. He assumed the MOE variation as a Gaussian random field simulated through the Nataf transformation [4]. Köhler [5] reported a model of the MOE lengthwise variability following the weak-zone approach proposed by Isaksson [6] for the bending moment capacity. Baño et al. [7] study the influence of the size and position of cylindrical knots on the load capacity of timber elements. Guindos et al. [8] assess the influence of different types of knots and the fiber deviations on the bending of wood elements. García et al. [9] analyze the effect of the presence of knots in the natural vibration frequencies and modes of timber beams. The knots simulations are added to the other mechanical and physical properties of the elements. The forced vibration problem in timber elements is addressed by Labonnote et al. [10]. The authors report a hysteretic damping model applied to Timoshenko timber beams. The shear and bending loss factors are numerically obtained through experimental damping measures and a fitting using a proposed analytical expression.

In the present work, the stochastic dynamic behavior of timber beams with uncertain properties is studied. In first place, the free vibrations case is stated. Then, the damped forced vibration is addressed. For the first problem, the differential equations are formulated starting from the Euler-Bernoulli (E-B) beam theory. The lengthwise variability of the MOE and the second moment of area account for the knots presence. For the first parameter, the local reduction of the MOE due to the grain deviation produced by the knot presence is considered. For the second moment of area the timber knot parameters are modeled via the Joint Probability Mass Function (JPMF) obtained with experimental data from visual survey of beams. For the second problem the governing differential equations are stated starting from the Timoshenko beam theory. The damping is introduced through the elastic and shear complex moduli. This material damping model provides an adequate description of the energy dissipation under the action of harmonic loads. Uncertainties in this problem are introduced in the elastic and shear moduli and the corresponding bending and shear loss factor, respectively. The governing equations are discretized and numerically approximated through the Finite Element Method (FEM). The Probability Density Functions (PDFs) of the natural frequencies and the Frequency Response Function (FRF) are numerically obtained through Monte Carlo Simulations...
2 MODAL ANALYSIS

In a first place, we present the study of the natural vibration frequencies and modes of a pinned-pinned sawn beam of Argentinian *Eucalyptus grandis* with knots according to the Euler-Bernoulli (E-B) beam theory. For the free vibration problem the well-known differential equation is the following:

\[
P a \frac{\partial^2 V(x, t)}{\partial t^2} + \frac{\partial^2}{\partial x^2} \left( E(x) I(x) \frac{\partial^2 V(x, t)}{\partial x^2} \right) = 0
\]  

(1)

where \( P \) is the material density per unit length; \( a \) is the beam cross-section; \( E(x) \) is the Modulus of Elasticity (MOE); \( I(x) \) is the second moment of area of the beam cross-section; \( V(x, t) \) is the transverse displacement; \( x \) is the position along the beam length and \( t \) is the time variable. Random variables of the stochastic model are denoted by capital letters.

2.1 Mechanical properties of the models

In this section, we briefly present the assumptions and the way in which the mechanical properties that appear in Eq. (1) are presented. For more details on the topics which will be presented in this section, the reader is referred to García *et al.* [9].

2.1.1 Timber knots geometric parameters

In order to simulate the timber knots, we define the joint PMF of the timber knot shape parameters within the beam cross-section, and the probability mass functions of the distance between timber knots and of its length in the direction parallel to the longitudinal axis of the beam. To find the joint PMF of the knots parameters, experimental data obtained from a visual survey of 25 sawn beams of *Eucalyptus grandis* of structural size with 180 timber knots were employed.

2.1.2 Random field of the Modulus of Elasticity (MOE)

The MOE values in this model are assigned in function to the knot ratio \( K \) of each knot or knots clusters. The knot ratio is calculated as the ratio of the largest dimension of the major knot or the sum of the individual dimensions of the knots that conform a cluster and the transverse dimension of the surface in which they appear. In the beam sections with \( K \) equal to zero (i.e. free-of-knots sections), the MOE values are assigned through the Nataf transformation [4]. In the beam sections with \( K \) different from zero (sections with knots), the MOE values are uncorrelated random variables assigned through their marginal PDF defined for each value of \( K \).

The correlation structure of the random field, when \( K \) is equal to zero, is described by means of the correlation function. We assume an exponential correlation function,
proposed by Czmoch [3] \( \rho_{ij} = \exp \left[ -2 \left( |x_{c}^{(j)} - x_{c}^{(i)}| / d \right) \right] \) where \( d \) is the correlation length, which measures the decay of the correlation function. In this work, the values considered for the correlation length of the MOE random field, are \( d = 1.34 \m (\text{named } d_1) \) and \( d \to \infty \) (named \( d_\infty \)). When \( d \to \infty \), the random field becomes fully correlated and it can be interpreted as a random variable in the limit.

### 2.1.3 Marginal PDF of the MOE

In order to determine the marginal PDF of the MOE, the Principle of Maximum Entropy PME [13] is applied. The measure of uncertainties of the continuous random variable \( X \) is defined by the following expression

\[
S(f_X) = -\int_D f_X(x) \ln(f_X(x)) \, dx
\]

in which \( f_X \) stands for the PDF of the random variable \( X \) and \( D \) is its domain. It is possible to demonstrate that the application of the principle under the constraints of positiveness and bounded second moment, leads to a gamma PDF. The PME conduces to this PDF due to the fact that the domain of the MOE is real and positive.

To find the parameters of the marginal PDF of the MOE, experimental data presented by Piter [1] are employed. These values were obtained by means of two point load bending tests, performed with 349 sawn beams of Argentinean \( Eucalyptus grandis \).

### 2.1.4 Mass density

In this work, the mass density in sawn beams of Argentinian \( Eucalyptus grandis \) is considered constant along the beam span and the lengthwise variability due to the knot presence is not taken into account. In the timber structural elements of this specie, the knots are frequently composed of material with similar density than the clean wood. Again, a gamma PDF is obtained from the application of the PME.

### 2.2 Numerical results

#### 2.2.1 Modal analysis

Numerical results of the first three natural frequencies \( F_n \) and modes \( \Phi_{F_n} \) \((n = 1 \cdots 3)\) are reported in this section. The equation of motion is discretized through the FEM with E-B beam elements with two nodes and degree of freedom per node with Hermitian shape functions [14]. The dimensional parameters of the timber knots are simulated with the inverse transform method [12]. Timber beams employed in the numerical simulation have the following dimensions: a length of 3 m and a nominal section of 50 mm x 150 mm.

A convergence study is shown in Fig. 1, where \( ns \) is the number of independent Monte Carlo Simulations (MCS) and \( E[F_1] \) is the mean value of the first natural frequency. The
adopted convergence criterion is the following: \(|E[F_{ns}^1] - E[F_{ns-200}^1]| \leq 0.1 \text{ Hz} \); where \(E[F_{ns}^1]\) is the mean value of the first natural frequency for a number of simulations \(ns\) and \(E[F_{ns-200}^1]\) is the mean value of the first natural frequency for a number of simulations \(ns - 200\). In this figure, the difference in \(E[F_1]\) for the two considered values of correlation length for the free-of-knots section of the beam can be observed. The probabilistic model for \(d_\infty\) shows higher values of \(E[F_1]\) than for the same model for \(d_1\), and the plots of the mean value variation with respect to \(ns\) are more separated. This study was carried out in

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{Convergence of \(E[F_i]\) for \(d_1\) and \(d_\infty\).}
\end{figure}

order to determine the appropriate number of simulations to attain a prescribed accuracy taken into account the results that will be shown in the work.

### 2.2.2 Validation of the stochastic model

In this section, a comparison between numerical and physical experiments through the analysis of the first natural frequency is presented. Experimental data, obtained by means of the test performed with 50 sawn beams of Argentinian *Eucalyptus grandis* are employed.

The comparison of the cumulative distribution function (CDF), of the first natural frequency found with numerical and experimental approaches is shown in Fig. 2. Numerical results correspond to two different values of the weak-zone length. The left plot was obtained with a weak-zone length equal to five times the major knot dimension, while the right plot corresponds to a weak-zone length equal to seven times the major knot dimension. Numerical results for the first case and \(d_1\) are acceptable and show a good prediction of the first natural frequency of the tested beams.

### 2.2.3 Natural frequencies and mode shapes

Figure 3 depicts the PDFs of the first three natural frequencies \(F_n\). The influence of the correlation length \(d\) adopted for the free-of-knots sections can be appreciated. The PDF of \(F_n\) have a same initial point. In this region, the beams have a larger amount
of knots and the lengthwise variation of the MOE in the sections without knots is not relevant. Towards the higher frequencies of the response range the values of \( d \) are more important.

![Figure 2: Comparison between numerical and experimental results CDF of \( F_1 \). Two different weak-zone lengths. Left plot: five times the major knot dimension; right plot: seven times the major knot dimension.](image)

![Figure 3: PDF of the first three natural vibration frequencies for \( d_1 \) and \( d_\infty \).](image)

The mean value and the standard deviation functions \( E[\Phi_{F_n}(x)] \) and \( \sigma[\Phi_{F_n}(x)] \) of the stochastic processes \( \Phi_{F_n}(x) \) are presented in Fig. 4. The mean value function is not modified by the values of the correlation length. On the other hand, a decrease in the value of \( d \) produces an increment in the standard deviation function \( \sigma[\Phi_{F_n}(x)] \). The inhomogeneity of the MOE values in the free-of-knots sections produces larger variations in the natural vibration modes for the second and third natural modes.

### 3 FORCED VIBRATIONS

The forced vibration of a Timoshenko timber beam is now presented. This beam formulation is selected for the study of the energy dissipation due to the bending and shear deformations of the beam. It is known that timber material exhibits a large ratio between the shear and bending elastic moduli of around 1/16 which usually induces a non-negligible
shear deformation. Following the work of Labonnote et al. [10] the longitudinal and shear loss factors, that account for the energy dissipation due to bending and the shear behavior of the beam, are employed. They are introduced in the equation of motion of the beam following the premises of the material damping approach also called complex stiffness (Clough [11]). This model of damping provides an energy dissipation not-dependent of the excitation frequency. The application of the Hamilton’s Principle leads directly to the equation of motion:

\[
[M] \ddot{x} + \frac{n_E}{w} [K_E] \dot{x} + \frac{n_G}{w} [K_G] \dot{x} + [K_E] x + [K_G] x = \{F_0\} e^{iwt} \tag{3}
\]

After replacing \(\{x\} = \{X_0\} e^{iwt}\) and its derivatives in Eq. (3), we obtain the following expression:

\[
(-w^2 [M] + [K_E] + [K_G]) + i (n_E [K_E] + n_G [K_G])) \{X_0\} = \{F_0\} \tag{4}
\]

As can be observed in Eq. (4), the damping is provided by the complex stiffness of the beam due to the bending and shear behavior, \(n_E [K_E]\) and \(n_G [K_G]\), respectively.

The equation of motion is discretized through the FEM with Timoshenko beam elements with two nodes and degrees of freedom per node with cubic and quadratic shape functions [14]. The main advantage of the Timoshenko formulation and the dissipation model herein presented is the possibility of discretizing the potential and dissipated energy in its components due to the bending and shear behavior of the beam. To accomplish this, the potential energy is obtained from \(U_B = \frac{1}{2} X'[K_E] X\) and \(U_S = \frac{1}{2} X'[K_G] X\) where \(X\) is the vector of the nodal displacements and the dissipated energy is obtained from \(W_B = 2\pi n_E U_B\) and \(W_S = 2\pi n_G U_S\) where \(B\) and \(S\) indicates the bending and shear components of the energy. The Frequency Response Function (FRF) of the vertical displacement at the free end of the cantilever timber beam is obtained through \(h(w) = \tilde{u}_L(w) / \tilde{f}_L(w)\) in which \(\tilde{u}_L\) and \(\tilde{f}_L\) are the Fourier transform of the vertical displacement and of the applied force at \(x = L\), respectively.

Figure 4: Mean value and the standard deviation functions \(E[\Phi_{Fn}(x)]\) and \(\sigma[\Phi_{Fn}(x)]\) of the stochastic processes \(\Phi_{Fn}(x)\).
3.1 Mechanical properties of the models

The MOE is modeled alternatively, as a random variable and as a random field through the Nataf Transformation with an exponential correlation function and gamma marginal PDF with mean value $\mu = 13.52$ GPa and standard deviation $\sigma = 3.6$ GPa. As a consequence of the relation between the bending and shear moduli, the last one is obtained from the ratio $G = E/16$ for its simulation as a random variable and as a random field. The mass density is modeled as a gamma random variable with mean value $\mu = 504.79$ kg/m$^3$ and standard deviation $\sigma = 34.59$ kg/m$^3$ assuming a constant value within the beam.

In order to determine the PDF of the random variables that represents the bending and shear loss factors, the PME is employed. We propose that the mean value of these random variables and their interval are known. For the latter, we suppose that the random variables can adopt any value between zero and twice the mean value. The application of the PME under the mentioned constrains leads to the truncated exponential PDF:

$$f(x) = \mathbb{1}_{[a,b]}(x) \exp(-\lambda_0 - \lambda_1 x)$$

3.2 Numerical results

3.2.1 Energy dissipation

The mean model employed in this study consists in a cantilever Eucalyptus grandis beam with a harmonic force applied at the free end with an amplitude of 500 N and a frequency of 10 Hz. The timber beam has the following dimensions: length $L = 2$ m; beam height $h = 0.1$ m and beam width $b = 0.05$ m. The mean material properties are: bending elastic modulus $E = 13.52$ GPa; shear modulus $G = E/16$; mass density $\rho = 504.79$ kg/m$^3$; bending loss factor $n_E = 0.0113$ and shear loss factor $n_G = 0.2769$. These values of bending and shear loss factor are presented by Labonnote et al. [10] for natural frequencies lower than 35 Hz.

The total potential energy is dominated by the bending contribution as can be seen in Fig. 5 left plot. However, in spite of this, it is interesting to see that a 11.16 % of the results show a dissipated shear energy larger than the bending one when $E$ and $G$ are modeled as random variables. On the other hand, when the material properties are modeled as random fields, this percentage increases to 12.20 %. The existence of samples with a shear dissipated energy larger than the bending one is due to the values that can assume the random variables $n_E$ and $n_G$ and its combinations with $E$ and $G$.

3.2.2 Frequency Response Function (FRF)

The FRF considering only the loss factors as stochastic variables of the probabilistic model is shown in Fig. 6. It can be observed that for lower frequencies and for the ones found out of the resonance zones, the influence of the random loss factors is low. The maximum and minimum values of $H(w)$, blue and red lines, are close to the FRF of
Figure 5: Potential and dissipated energy obtained for the probabilistic model with the random variables $E$, $G$, $\rho$, $n_E$ and $n_G$.

the mean model. In the frequencies zones near to the second and third resonance and anti-resonance, the influence of the stochastic loss factors increases.

Figure 6: FRF for the probabilistic model with the random variables $n_E$ and $n_G$.

In Fig. 7, the FRF for the stochastic model considering $E$ and $G$ as random variables is shown. The system responses considering alternatively the loss factors as deterministic or stochastic variables, are presented. The influence of the random loss factors is observed for the frequencies in the resonances zones. Outside of these regions, the differences between the approaches employed for modeling the loss factors, stochastic or deterministic, are negligible.

The FRF for the stochastic model considering $E(x)$ and $G(x)$ as random fields is depicted in Fig. 8. The loss factors are considered in the same manner than in the previous case and present the same influence. The difference between the results of the Figures 7 and 8 can be appreciated in the resonance frequencies bandwidth. In the first figure, the resonance frequency range is larger than in the second one. This is an effect produced due to the lengthwise variability of $E(x)$ and $G(x)$. 
Figure 7: FRF obtained for \( E \) and \( G \) with mean and random damping.

Figure 8: FRF obtained for \( E(x) \) and \( G(x) \) with mean and random damping.

Fig. 9 shows the histograms of the FRF for the first three resonance zones of Fig. 7. The upper plots show the results obtained when the loss factors are considered deterministic with their mean values. Meanwhile, the lower plots show the results obtained for random loss factors. As can be viewed, the shape and the range of the histograms are modified according to the manner of considering the loss factor, deterministic or stochastic. Similar histograms have been obtained when the lengthwise variability of the materials properties was considered.

4 CONCLUSIONS

A stochastic dynamic study of a timber beam was presented. Two probabilistic models were studied. The first, employed for the modal analysis, starts from the weak-zone representation. This model introduces the presence of knots in the sectional parameters, the length of the weak zone, the choice of the parameters of the marginal PDF of the MOE depending of the section knot ratio (\( K \)) and the lengthwise variability of the MOE in the free-of-knots section. The second one, employed for the forced vibration problem,
Diego A. García, Marta B. Rosales and Rubens Sampaio

Figure 9: Histograms of $H(w)$ for frequencies in the first three resonance zones. Upper plots: deterministic loss factors; lower plots: random loss factors.

is a traditional timber model that only considers the lengthwise variability of the MOE.

For the first problem, the PDFs of the first three natural frequencies were obtained through Monte Carlo Simulations (MCS). The influence of the correlation length in the free-of-knots sections was evaluated. Validation of the numerical results of the first natural frequency with experimental values was also presented. Thus, the length of the weak-zones was calibrated and it is shown that numerical simulations with the MOE in the free-of-knots section represented by a random field are closer to the experimental outcome. A stochastic process of the normalized mode shapes parametrized by the position is obtained. The mean value and the standard deviation functions of the stochastic processes were presented. The latter increase their values from the first to the third natural frequencies.

For the second problem, the complex stiffness damping was introduced through the bending and shear loss factors. Then, the energy dissipation was studied within the range of frequencies where this damping model is applicable. The potential and dissipated energy due to bending and shear behavior of the beam were obtained. Although the fact that the total potential energy is dominated by the bending contribution, numerical results show a dissipated shear energy larger than the bending one in some samples. The FRFs for different configurations of the probabilistic model were presented. They shown the influence of the bending and shear elastic modules represented as random variables and as random fields. The FRFs and the histograms show that as the frequency increases, the predictability decreases. Despite the fact that the damping model herein presented is not applicable for natural frequencies higher than 35 Hz, interesting results regarding with the displacements of the structural system for higher frequencies were found.

5 ACKNOWLEDGEMENTS

The authors are grateful for the financial support of CONICET, SGCyT-UNS, MINCyT from Argentina and CAPES, CNPq, and FAPERJ from Brazil. The experimental data provided by J.C. Piter and co-workers from FRCU-UTN (Argentina) is greatly acknowledge.
REFERENCES


PARAMETRIC PROBABILISTIC ANALYSIS IN THE
UNCERTAIN DYNAMICS OF MAGNETO-ELECTRO-
ELASTIC BEAMS RESTING ON ELASTIC FOUNDATIONS

MARCELO T. PIOVAN\textsuperscript{1}, JOSÉ F. OLMEDO \textsuperscript{2} AND RUBENS SAMPAIO\textsuperscript{3}

\textsuperscript{1}Centro de Investigaciones de Mecánica Teórica y Aplicada, Universidad Tecnológica Nacional FRBB and CONICET, 11 de Abril 461, B8000LMI, Bahía Blanca. Argentina
\texttt{mpiovan@frbb.utn.edu.ar}

\textsuperscript{2}Departamento de Ciencias de la Energía y Mecánica, ESPE-Sangolquí/SENESCYT Av. Rumiñahui s/n Sangolquí, Ecuador.
\texttt{jfolmedo@espe.edu.ec/mpiovan@espe.edu.ec}

\textsuperscript{3}Departament of Mechanical Engineering, Pontifícia Universidade Católica do Rio de Janeiro Marquês de São Vicente 225, Rio de Janeiro RJ-22453-90
\texttt{rsampaio@puc-rio.br}

Key words: Uncertain dynamics, MEE beams, elastic foundations

Abstract. In this paper, the authors are concerned with uncertainty quantification of the dynamic behavior of magneto-electro-elastic (MEE) beams (both straight and curved) mounted on elastic foundations. The MEE beams are a type of composite structures that can be employed as imbedded subsystem in high performance aero spatial structures, control of motion and attenuation of vibrations, energy harvesting, etc. Although most of the research work related to the mechanics of MEE structures was done for dynamics and statics, it is remarkable the scarcity of papers analyzing the uncertainty propagation in the dynamics of a MEE structure, considering that many models have uncertainties related to their parameters (loads and/or material properties) and/or the hypotheses invoked to develop the model, among others. In order to perform studies about the uncertainty quantification of MEE beams, a beam model for curved beams is derived and employed as a mean basis or deterministic approach to the studies on stochastic modeling. Then, a probabilistic model is constructed, on the basis of the finite element formulation of the deterministic approach, by adopting random variables for the uncertain parameters selected. The probability density functions of the random variables are derived appealing to the Maximum Entropy Principle. Once the probabilistic model is constructed, the Monte Carlo method is used to perform simulations with independent random variables as uncertain parameters. Studies with several types of constitutive proportions are carried out in order to characterize the Magneto-elastic and/or piezoelectric coupling and response of the aforementioned MEE beams.
1 INTRODUCTION

The MEE material are a kind of smart composites exhibiting various coupling effects that can be of useful in many high-tech structural applications. That is why many investigation on the mechanics of MEE structures have received considerable attention of the research community since the last 10 years.

Especial composite materials consisting of piezoelectric and piezomagnetic components are used in smart structures such as sensors, actuators, hydrophones, etc. The smart structures provide remarkable capabilities of sensing and reacting to external actions and/or disturbances, also satisfying reliability, light weight and the appropriate performances demanded in high-tech structural applications [1, 2].

In the last 10/15 years many researchers developed new theories and technical models for studying the mechanical response (statics, dynamics, instability, etc) of the so-called magneto electro elastic (MEE) structures. An interesting variety of models of MEE structures has been introduced principally for piezoelectric and piezomagnetic plates and shells [2, 3]. In these articles the static behavior of multilayered MEE strips and plates was analyzed by means of 3D formulation and by subjecting the specimens to simpler loads, that is, to sinusoidally distributed magnetic, electric and mechanic loads. Wu and Lu [4] and Tsai et al. [5] among others carried out extensive studies in the dynamics responses of shells and plates appealing to 3D formulations. There was also a research of MEE shells with simply or doubly curved profile [6].

According to a thorough bibliographical review, it is remarkable that there are quite few articles related to the study of MEE slender structures in the context of a beam model or theory. The papers of Milazzo and coworkers [7, 8] are, apparently, the very first in which studies about the dynamics of MEE beams (for enhanced Bernoulli-Euler and/or Timoshenko theories) have been carried out. Besides these works are quite recent. However to the best of authors’ knowledge, articles related to static/dynamic behavior of MEE curved beams are apparently absent.

On the other hand it should be recognized that many aspects related to the construction of MEE materials and/or structures are connected with a variable source of uncertainty that can substantially alter the response of the structure. Possible sources of uncertainty can be found in material properties, boundary conditions, loads [10], the hypotheses of model or the model itself [11], etc. In order to characterize the uncertain response in dynamics of structures there is a bunch of alternatives that can be collected in two master sets: parametric probabilistic approach (PPA) [10] and non-parametric probabilistic approach (NPPA) [11]. In the first case the source of uncertainty are the parameters of the model in the second case the model as a whole. In the PPA the uncertain parameters are associated to a random variable whose probability density function (PDF) is defined according to given information about them (mean values, standard deviation, bounds, etc.).

Thus, the scope of this research is directed toward offering some contributions in the
mechanics/dynamics of curved MEE beams and especially to quantify the propagation of the uncertain in the dynamic response of curved and also straight MEE beams. In this context, the present article is arranged according to the following scheme: As first step the hypotheses of the constitutive model are enunciated and the deterministic structural model is presented, then an equivalent MEE curved beam model is constructed on the basis of curved beam models previously developed by the first author [13] which are conceived in the context of first order shear theories. A finite element formulation is proposed and then employed to carry out calculations of the deterministic model. Subsequently, the probabilistic model is constructed employing the previous finite element formulation in which the random variables are incorporated. The PDF of the random variables (some elastic, electric properties, elastic foundations, etc) are deduced by employing the maximum entropy principle [9] subjected to given known information such as expected values and/or COV. Then the Monte Carlo method is employed to simulate realizations, a the statistical analysis is done and the results presented in the form of frequency response functions or other graphics of statistical interest.

2 FORMULATION OF THE DETERMINISTIC MODEL

2.1 Basic hypotheses of the structural model

The MEE structure of this paper consists of a thin curved strip as one can see in Fig. 1 with the reference system located in the geometric centroid \( C \). The curved beam has a circumferential length of \( L = R\beta \), a radial thickness of \( h \), width of \( b \) and a constant radius of curvature \( R \).

![Diagram of the Curved MEE beam](image)

The deterministic model for this study is based on the following assumptions: (a) The motion of the curved beam is constrained in the curvature plane (XY), (b) Shear flexibility is considered, (c) the material is supposed to be poled in the radial direction and it consists of a mixture in given proportions of \( BaTiO_3 \) and \( CoFe_2O_4 \), (d) the electric and magnetic fields are determined through their corresponding potentials which are prescribed on the cylindrical surfaces (i.e. \( y = R \pm h/2 \), or \( y = R_i, y = R_o \)); (e) The radial components of the electric and magnetic fields are substantially greater than the
circumferential components \((E_x \ll E_Y H_x \ll H_y)\), (f) A generic elastic foundation (characterized with spring coefficients) is assumed, (g) the structural damping is considered as "a posteriori" addition in the finite element formulation.

Employing the hypotheses, the displacement field can be derived [13] as:

\[
\begin{align*}
    u_x(x, y, t) &= u_{xc} - y \left( \theta_z - \frac{u_{xc}}{R} \right) \\
    u_y(x, y, t) &= u_{yc}
\end{align*}
\]

where \(u_{xc}\) and \(u_{yc}\) are the circumferential and radial displacements of the reference point \(C\) whereas \(\theta_z\) rotation parameter. The representative strain components can be written in the following form [13]:

\[
\begin{align*}
    \varepsilon_{xx} &= (\varepsilon_{D1} - y\varepsilon_{D2}) F \\
    \gamma_{xy} &= \varepsilon_{D3} F
\end{align*}
\]

where:

\[
\begin{align*}
    \varepsilon_{D1} &= u_{xc}' + \frac{u_{yc}'}{R}, \varepsilon_{D2} = \theta_z' - \frac{u_{xc}'}{R}, \varepsilon_{D3} = u_{yc}' - \theta_z, F = \frac{R}{R + y}
\end{align*}
\]

In Eq. 3, the apostrophes represent derivatives with respect to the spatial variable \(x\). Moreover, \(\varepsilon_{D1}\) can be interpreted as the generalized circumferential strain, \(\varepsilon_{D2}\) as the generalized bending curvature and \(\varepsilon_{D3}\) as the generalized shear strain.

The mechanical equilibrium equations of the curved MEE beam supported on the elastic foundation can be written in the following form:

\[
\begin{align*}
    -Q_x' + \frac{M_z'}{R} + k_1 u_{xc} + \mathcal{M}_1 \left( \ddot{u}_{xc}, \ddot{u}_{yc}, \dot{\theta}_z \right) - \mathcal{P}_1 (x) &= 0 \\
    -Q_y' + \frac{M_z}{R} + k_2 u_{yc} + \mathcal{M}_2 \left( \ddot{u}_{xc}, \ddot{u}_{yc}, \dot{\theta}_z \right) - \mathcal{P}_2 (x) &= 0 \\
    -M_z' - Q_y + k_3 \theta_z + \mathcal{M}_3 \left( \ddot{u}_{xc}, \ddot{u}_{yc}, \dot{\theta}_z \right) - \mathcal{P}_3 (x) &= 0
\end{align*}
\]

with the corresponding boundary conditions.

In the previous expressions, \(Q_x\) is the axial force, \(Q_y\) is the shear force \(M_z\) is the bending moment, \(\mathcal{P}_i, i = 1, 2, 3\) represent distributed forces and moments, whereas \(\mathcal{M}_i, i = 1, 2, 3\). All these entities are defined as:

\[
\{Q_x, Q_y, M_z\} = \int_A \{\sigma_{xx}, \sigma_{xy}, -y\sigma_{xx}\} \, dA
\]

\[
\begin{bmatrix}
    \mathcal{M}_1 \\
    \mathcal{M}_2 \\
    \mathcal{M}_3
\end{bmatrix}
= \begin{bmatrix}
    J_{11}^p & 0 & J_{13}^p \\
    0 & J_{22}^p & 0 \\
    J_{13}^p & 0 & J_{33}^p
\end{bmatrix}
\begin{bmatrix}
    \ddot{u}_{xc} \\
    \ddot{u}_{yc} \\
    \dot{\theta}_z
\end{bmatrix}
\]

In Eq. (6) \(J_{ik}^p, i, k \rightarrow 1, 2, 3\) are inertia constants that are described extensively in Ref. [14].
2.2 Deduction of Potentials and Constitutive equations

The constitutive equations of a magneto-electro-elastic solid under the hypothesis of plane stress - assuming $\sigma_{yy} << \sigma_{xx}$ and employing the hypothesis (d) - can be written in the following matrix form [15, 7, 8]:

\[
\begin{bmatrix}
\sigma_{xx} \\
\sigma_{xy} \\
D_x \\
D_y \\
B_x \\
B_y
\end{bmatrix} =
\begin{bmatrix}
c_{11} & 0 & -c_{21}^* & -q_{21}^* \\
0 & c_{66} & 0 & 0 \\
0 & 0 & e_{16}^* & 0 \\
e_{21}^* & 0 & \eta_{22} & d_{22} \\
0 & q_{16}^* & 0 & 0 \\
q_{21}^* & 0 & d_{22} & \mu_{22}
\end{bmatrix}
\begin{bmatrix}
\varepsilon_{xx} \\
\gamma_{xy} \\
E_y \\
H_y
\end{bmatrix}
\tag{7}
\]

and $\sigma_{ij}$ are the stresses, $\varepsilon_{ij}$ and $\gamma_{xy}$ are the axial and shear strain components; $D_i$ and $B_i$ are the components of the electric displacement and the magnetic flux, respectively; $E_i$ and $H_i$ are the components of the electric field and the magnetic field, respectively; $c_{ik}^*$ are elastic coefficients; $\eta_{ij}$ are the dielectric coefficients; $\mu_{ij}$ are the magnetic permeability coefficients; $e_{ij}^*$ are the piezoelectric coefficients; $q_{ij}^*$ are the piezomagnetic coefficients and $d_{ij}$ are the magnetoelectric coefficients.

It is interesting to note that from Eq. (7), the electric displacement and magnetic flux in x-direction ($D_x$ and $B_x$, respectively) depend only on the shear strain $\gamma_{xy}$. These constitutive expressions allow the calculation of the electric and the magnetic potentials in explicit form (see reference [7] for detailed issues).

The electric and magnetic fields are defined in terms of the electric and magnetic potentials (i.e. $\vec{E} = -\vec{\nabla}\Phi$ and $\vec{H} = -\vec{\nabla}\Psi$, with $\vec{\nabla}$ as the gradient operator). The equilibrium equations of electrostatics and magnetostatics are [15]:

\[
\vec{\nabla} \cdot \vec{D} = 0 \\
\vec{\nabla} \cdot \vec{B} = 0
\tag{8}
\]

where $\vec{D} = \{D_x, D_y, D_z\}$, $\vec{B} = \{B_x, B_y, B_z\}$, and $\vec{\nabla}$ is the divergence operator (in curvilinear coordinates).

Now, employing Eq. (7) together with Eq. (8) and after some algebraic handling it is possible to arrive to:

\[
\Phi = \mathcal{D}_1 \Phi_0 + \mathcal{D}_2 \Phi_1 + \mathcal{D}_3 \varepsilon_D' + \mathcal{D}_4 \varepsilon_{D2} \\
\Psi = \mathcal{B}_1 \Psi_0 + \mathcal{B}_2 \Psi_1 + \mathcal{B}_3 \varepsilon_{D3} + \mathcal{B}_4 \varepsilon_{D2}
\tag{9}
\]

where the coefficients $\mathcal{D}_i$, $\mathcal{B}_i$, $i = 1, ..., 4$ defined extensively in Ref [14].

According to the definition of the internal forces given in Eq. (5) and employing Eq. (7) it is possible to derive the following expression of the internal forces in terms of strain components and electro/magnetic potentials:

\[
\mathbf{Q} = \mathbf{M}_{CA} \vec{\varepsilon}_D + \mathbf{M}_{CB} \vec{\varepsilon}_D' + \mathbf{M}_{CB} \mathbf{P}_{\Phi \Psi}
\tag{10}
\]
where:

\[
\bar{Q} = \{Q_x, M_z, Q_y\}^T, \bar{\epsilon}_D = \{\epsilon_{D1}, \epsilon_{D2}, \epsilon_{D3}\}^T
\]

\[
\bar{P}_{\Phi \Psi} = \{\Phi_0, \Phi_1, \Psi_0, \Psi_1\}^T
\]

\[
M_{CA} = \begin{bmatrix}
J_{X1} & J_{X2} & 0 \\
J_{M1} & J_{M2} & 0 \\
0 & 0 & J_{Y3}
\end{bmatrix},
\]

\[
M_{CB} = \begin{bmatrix}
0 & 0 & J_{X4} \\
0 & 0 & J_{M4}
\end{bmatrix},
\]

\[
M_{CE} = \begin{bmatrix}
J_{X5} & J_{X6} & J_{X7} & J_{X8} \\
J_{M5} & J_{M6} & J_{M7} & J_{M8}
\end{bmatrix}
\]

besides \(J_{Xi}, J_{Mi}, J_{Yi}, i = 1, \ldots, 8\) are constants that depend on the elastic, electric and magnetic properties (extensive expressions can be followed in Ref. \[14\]).

### 2.3 Weak formulation and finite element approach of MEE curved model

Substituting Eq. (10) in Eq. (4) and incorporating the boundary conditions given in Eq. (4) it is possible to write the following weak form of the equilibrium equations of the MEE curved beam model:

\[
\int_L \left[ Q_x \delta \epsilon_{D1} + Q_y \delta \epsilon_{D2} + M_z \delta \epsilon_{D3} + \mathcal{M}_1 \delta u_{xc} + \mathcal{M}_2 \delta u_{yc} + \mathcal{M}_3 \delta \theta_z \right] \, dx + \\
\int_L \left[ (k_1 u_{xc} - P_1) \delta u_{xc} + (k_2 u_{yc} - P_2) \delta u_{yc} - (k_3 \theta_z + P_3) \delta \theta_z \right] + \\
- \left. \left[ \left( Q_x - \frac{M_z}{R} \right) \delta u_{xc} + Q_y \delta u_{yc} - M_z \delta \theta_z \right] \right|_{x=0}^{L=0} = 0
\]

(12)

The weak variational formulation (12) represents the equilibrium mechanics of a MEE curved beam. The solution of this formulation for given problems (dynamics and/or statics) can be tackled by means of many approaches: Generalized direct variational approaches, finite elements approaches among many other alternatives.

A Finite Element formulation can be derived through discretization of Eq. (12). The discretization is carried out using isoparametric elements with five nodes and shape functions of quartic order.

Each kinematic variable \(u_{xc} \equiv U_1, u_{yc} \equiv U_2\) and \(\theta_z \equiv U_3\), can be interpolated within the element by means of following compact form:

\[
U_i = F_i \bar{U}_e, \quad i = 1, \ldots, 3
\]

(13)

where the matrix \(F [3 \times 15]\) collects the shape functions \(f_j (\bar{x}), j = 1, \ldots, 5\), for the isoparametric elements of quartic order. Each variable \(U_i, i = 1, \ldots, 3\) is interpolated with the same shape functions in \(F\). \(\bar{U}_e\) is the vector of kinematic nodal variables.
Now substituting Eq. (13) in Eqs. (3) and (10) and then in Eq. (12) it is possible to derive the following equation in the elementary domain:

\[ \delta U_e^T \left[ \int F^T D_{D1}^T M_{CA} D_{D1} F e d\xi \right] + \delta U_e^T \left[ \int F^T D_{D1}^T M_{CB} D_{D1} F e d\xi \right] + \delta U_e^T \left[ \int D_{D1}^T M_{CE} \Phi \Psi F e d\xi \right] + \delta U_e^T \left[ \int D_{D1}^T M_{CI} F e d\xi \right] + \delta U_e^T \left[ \int D_{D1}^T \bar{P} F e d\xi \right] = 0 \]

where \( l_e \) is the length of the element, \( \xi = x/l_e \) is the internal coordinate, \( M_{CA} \), \( M_{CB} \), \( M_{CE} \) and \( \Phi \Psi \) are defined in Eq. (11), \( M_{CI} \) is defined in Eq. (6), \( \bar{P} \) is the vector of distributed forces, \( \bar{F}_{BC} \) is a vector of prescribed forces at the ends of the element, the matrix \( M_{CF} = \text{diag}(k_1, k_2, k_3) \) is the matrix of elastic foundation constants and \( D_{D1} \) is a differential operator.

Employing the usual assembly procedures of the finite element method with the matrix structure of Eq. (14) and incorporating the "a posteriori" proportional damping \( C_{SD} = \eta_1 M + \eta_2 K \) one gets:

\[ K \bar{W} + C_{SD} \dot{\bar{W}} + M \ddot{\bar{W}} = \bar{F}_M - \bar{F}_{em}. \]

where \( K \) and \( M \) are the global matrices of elastic stiffness and mass, respectively; whereas \( \bar{W} \), \( \dot{\bar{W}} \) and \( \ddot{\bar{W}} \) are the global vectors of nodal displacements, velocities and accelerations, finally \( \bar{F}_M \) and \( \bar{F}_{em} \) are the global vectors of forces due to mechanical components strictly and forces due to electromagnetic loads applied in boundaries and/or the curved beam span, respectively.

From the previous Eq. (15) two possible issues are analyzed in the present investigation: transient analysis in the frequency domain and eigenvalues. Thus, the response in the frequency domain (after a Fourier transformation) discretized model given in Eq. (15) is written as:

\[ \hat{W}(\omega) = [-\omega^2 M + i\omega C_{SD} + K]^{-1} \hat{F}(\omega), \]

where \( \hat{W} \) and \( \hat{F} \) are the Fourier transform of the displacement vector and force vector, respectively; whereas \( \omega \) is the circular frequency measured in [rad/sec]. Moreover, Eq. (15) can be reduced to calculate the dynamic eigenvalues (or natural frequencies) by neglecting damping and all applied forces and by assuming harmonic motion.
3 CONSTRUCTION OF THE PROBABILISTIC MODEL

In order to construct the probabilistic model, the previously developed deterministic model, in its finite element formulation, is employed as a mean expected response. The Maximum Entropy Principle (MEP) is used to derive the probability density functions (PDF) of the random variables associated with the uncertain parameters [9]. This particular is quite sensitive in stochastic analysis and PDF’s should be deduced according to the given information (normally and sensitively scarce) about the uncertain parameters. The deterministic model developed in the previous sections has many parameters that can be uncertain, however the most relevant could be the material properties and the spring constants of the elastic foundation.

In the present problem random variables \( V_i, i = 1, 2, 3 \ldots 13 \) are introduced such that they represent the equivalent material properties \( (i = 1, \ldots, 10) \) and the elastic foundation parameters \( (i = 11, 12, 13) \). The expected value of the random variables is known and it has the nominal value of the deterministic model, i.e.: \( \mathcal{E}\{V_i\} = V_i, i = 1, 2, 3 \ldots 13 \); moreover the random variables have bounded supports whose upper and lower limits can be represented in terms of given information (standard deviation or coefficient of variation). Provided that there is no information about the correlation or dependency of material properties, random variables \( V_i, i = 1, \ldots, 13 \), according to MEP, are assumed independent and non correlated. Consequently, taking into account the previous context, the PDF’s of the random variables can be written as:

\[
p_{V_i}(v_i) = \mathcal{S}[L_{V_i}, U_{V_i}](v_i) \frac{1}{2\sqrt{3}V_i\delta V_i}, \quad i = 1, \ldots, 13
\]

(17)

where \( \mathcal{S}[L_{V_i}, U_{V_i}](v_i) \) is the generic support function, whereas \( L_{V_i} \) and \( U_{V_i} \) are the lower and upper bounds of the random variable \( V_i \). \( V_i \) is the expect value (or deterministic value) of the \( i^{th} \) random variable, whereas \( \delta V_i \) is its coefficient of variation. The Matlab function \texttt{unifrnd(}\( V_i (1 - \delta V_i \sqrt{3}), V_i (1 + \delta V_i \sqrt{3}) \texttt{)}\) can be used to generate realizations of random variables \( V_i, i = 1, \ldots, 13 \). Then, using Eq. (17) in the construction of the matrices of finite element formulation given in Eq. (16), the stochastic finite element model can be written as:

\[
\hat{W}(\omega) = [-\omega^2\mathbf{M} + i\omega\mathbf{C}_{RD} + \mathbf{K}]^{-1}\hat{F}(\omega).
\]

(18)

Notice that in Eq. (18) the math-blackboard typeface is employed to indicate stochastic entities.

The Monte Carlo method is used to simulate the stochastic dynamics, which implies the calculation of a deterministic system for each independent realization of random variables \( V_i, i = 1, 2, \ldots, 13 \). The convergence of the stochastic response \( \hat{W} \) can be calculated with the following expression:
\[
\text{conv} (N_{MS}) = \sqrt{\frac{1}{N_{MS}} \sum_{j=1}^{N_{MS}} \int_{\Omega} \| \hat{W}_j (\omega) - \hat{W} (\omega) \| ^2 d\omega},
\]

where \( N_{MS} \) is the number of Monte Carlo samplings and \( \Omega \) is the frequency band of analysis. Clearly, \( \hat{W} \) is the response of the stochastic model and \( \hat{W} \) the response of the mean model or deterministic model.

4 COMPUTATIONAL STUDIES

4.1 Preliminary validations and comparative studies

In this section a comparison and validation of the deterministic MEE curved beam model with respect to other approaches is performed. The first example corresponds to a comparison of the present MEE curved beam model reduced to the case of straight beam (i.e. \( R \to \infty \)), with respect to the beam model of Milazzo et al [7] and the 2D finite element approach of Annigeri et al. [15]. The material properties of the MEE composite with 40\% \text{BaTiO}_3 and 60\% \text{CoFe}_2\text{O}_4 are taken from the work of Annigeri et al. [15] and the equivalent constants for the beam formulation are: \( c_{11}^* = 120.67 \text{ GPa}, \) \( c_{66}^* = 45 \text{ GPa}, \) \( e_{21}^* = 6.52 \text{ C/m}^2, \) \( e_{16}^* = 0, \) \( q_{21}^* = 32.66 \text{ N/(Am)}, \) \( q_{16}^* = 180 \text{ N/(Am)}, \) \( \eta_{22} = -8.97 \times 10^{-9} \text{ Ns}/(CV), \) \( d_{22} = 8.85 \times 10^{-9} \text{ F/m}, \) \( \mu_{22} = 7.54 \times 10^{-3} \text{ Ns}^2/C^2 \) and \( \rho = 5600 \text{ Kg/m}^3. \) These values are nearly the same of those calculated by Milazzo et al. [7] with a difference in percentage no higher than 0.7\%. The dimensions of the beam are \( L = 0.3 \text{ m}, h = 0.02 \text{ m} \) and \( b = 1.0 \text{ m} \) (in order to compare with the 2D FEM approach [15]). The boundary conditions can be clamped-free or doubly clamped.

In Table 1 it is possible to see the comparison of the first 5 natural frequencies calculated with the present model and the models derived by Milazzo et al. [7] and Annigeri et al. [15]. As it can be seen the present model can recover both extensional and bending modes with difference in percentage no higher than 1.0 \% with respect to Annigeri’s results [15] associated to a 2D FEM approach.

In Table 2 the first 5 frequencies of a curved MEE beam are shown and a comparison between various approaches and effects is also presented. The curved beam is clamped-free and has the same length, height and width of the previous example. In fact the present 1D model is compared with a 2D approach (performed under a plane-stress state in a general purpose finite element platform). Two alternatives in the 1D model are evaluated: considering full electromagnetic coupling in beam coefficients (i.e. in \( J_{Xi} \) and \( J_{Mi}, i = 1, ..., 4 \)) or neglecting that effect by imposing \( c_{21}^* = 0, e_{16}^* = 0, q_{21}^* = 0, q_{16}^* = 0. \) It is possible to note that 2D results are nearly compatible with the case where the full coupling is neglected with a difference in percentage lower than 1.0 \%. On the other hand, in the case of full MEE coupling in the calculation of beam sectional properties it is seen an increase of 3.0 to 4.0 \% in the frequencies when compared with the previous two cases.
Table 1: Natural Frequencies: comparison of the present model reduced to the case of straight beam

<table>
<thead>
<tr>
<th>Bounds</th>
<th>Mode Type</th>
<th>Present (Freq)</th>
<th>Annigeri (Freq)</th>
<th>Milazzo (Freq)</th>
<th>Clamped – Free</th>
<th>Clamped – Clamped</th>
</tr>
</thead>
<tbody>
<tr>
<td>clamp</td>
<td>Fle</td>
<td>169</td>
<td>170</td>
<td>169</td>
<td>Fle</td>
<td>1052</td>
</tr>
<tr>
<td>free</td>
<td>Fle</td>
<td>1041</td>
<td>1044</td>
<td>1043</td>
<td>Fle</td>
<td>2805</td>
</tr>
<tr>
<td>clamp</td>
<td>Fle</td>
<td>2833</td>
<td>2835</td>
<td>2831</td>
<td>Fle</td>
<td>5277</td>
</tr>
<tr>
<td>free</td>
<td>Ext</td>
<td>3867</td>
<td>3902</td>
<td>–</td>
<td>Ext</td>
<td>7734</td>
</tr>
<tr>
<td>clamp</td>
<td>Fle</td>
<td>5341</td>
<td>5337</td>
<td>5323</td>
<td>Ext</td>
<td>8321</td>
</tr>
</tbody>
</table>

Table 2: Natural Frequencies of a clamped-free curved MEE beam (considering and neglecting MEE coupling): comparison of the present model with 2D FEM approach (without piezoelectric effect)

<table>
<thead>
<tr>
<th>Relations</th>
<th>L/R = 1.0</th>
<th>L/R = 0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mode</td>
<td>2D FEM w/o coup</td>
<td>2D FEM w/o coup</td>
</tr>
<tr>
<td></td>
<td>2D FEM with coup</td>
<td>2D FEM with coup</td>
</tr>
<tr>
<td>1</td>
<td>170 169</td>
<td>170 167</td>
</tr>
<tr>
<td>2</td>
<td>921 921</td>
<td>938 1010</td>
</tr>
<tr>
<td>3</td>
<td>2629 2636</td>
<td>2682 2754</td>
</tr>
<tr>
<td>4</td>
<td>4358 4349</td>
<td>4363 3928</td>
</tr>
<tr>
<td>5</td>
<td>5292 5318</td>
<td>5398 5216</td>
</tr>
</tbody>
</table>

4.2 Uncertainties in the dynamic response

In this section a study about the propagation of uncertainty in the frequency response function (FRF) associated to several parameters is carried out. The example for the analysis is a clamped-free curved beam with the same dimensions of the precedent paragraph and \( L/R = 0.3 \). The FRF measure is characterized by \( H_F = u(L)/F(L) \) where \( F(L) \) is a unitary shock force [10].

In Fig. 2 it is shown the propagation of uncertainty associated to several parameters of the structure. Effectively, in Fig. 2(a) one can see how sensitive is the FRF to the variation of magneto-electric properties identified with a coefficient of variation \( \delta_{em} = 0.55 \) which is a little bit lower than the possible limit of the selected random variable (i.e. \( \delta_{em} < \sqrt{3} \), that identifies a 2\(^{nd}\) order random variable). On the other hand, a nearly extreme variation in the coupled piezo-electric/piezo magnetic parameters, identified by \( \delta_{pme} = 0.55 \), is not so sensitive in comparison to a standard variability, that is related to a variation of the order of \( \delta_{all} = \delta_{V_t} = 0.05 \) in all material parameters. In Fig. 2(b) it is possible to see effect of extreme variation in the elastic foundation properties that relies in a non sensitive uncertainty propagation.

5 CONCLUSIONS

In the present paper a new model for MEE curved beams has been presented. The model has been implemented in the context of finite element methodologies and employed
Figure 2: Uncertainty propagation due to several parameters. (a) In material properties, (b) In the elastic foundation.

to calculate natural frequencies of curved MEE beams. The model can reproduce results of straight beams with or without electromagnetic coupling. Also the curved beam model compares well with 2D approaches when the piezoelectric effect is neglected. Some preliminary studies for quantification of uncertainty due to some parameters of model have been carried out. The propagation of uncertainty due to material properties proved to be substantially more sensitive than the one related to the elastic foundation. Although in the last case it could be analyzed with a stochastic field, it should be the matter for future research based on the present one.

6 ACKNOWLEDGMENTS

The support of SCyT of Universidad Tecnológica Nacional (proyect: 2194TC) and Conicet (Argentina), SENESCYT under the Prometeo Program (Ecuador) and CNPq (Brazil) is kindly recognized.

REFERENCES


DYNAMIC ANALYSIS OF A GUYED MAST WITH UNCERTAINTIES ON THE STIFFNESS AND THE NOMINAL WIND VELOCITY

JORGE S. BALLABEN¹, MARTA B. ROSALES¹ RUBENS SAMPAIO²

¹ Department of Engineering - Universidad Nacional del Sur
Alem 1253, 8000, Bahía Blanca, Argentina
CONICET, Bahía Blanca, Argentina
jorgeballaben@gmail.com mrosales@criba.edu.ar

²Department of Mechanical Engineering, Pontificia Universidade Católica
Rua Marquês de São Vicente, 225, Rio de Janeiro, Brazil
rsampaio@puc-rio.br

Key words: Guyed Mast, Wind Loads, Non-Linear Dynamics, Uncertainty Quantification

Abstract. Nowadays the use of guyed structures as support for telecommunication antennas (radio, TV, internet, cell signals) is a practice widely adopted. Nevertheless, detailed studies of the dynamic response of the structural system under natural actions, as wind or earthquake are not frequent, despite the potential of adverse impact, particularly in the quality of the transmission. In this work, the dynamic response of a guyed mast is analyzed under wind load action. The simplified structural model consists of a beam-column model accounting for the second order effect due to axial loads and one guy. The guy is represented by an extensible cable governed by nonlinear equations. The wind load is calculated from the wind velocity field, by means of the Spectral Representation Method. The resulting governing system is discretized using finite elements and a reduced order model is afterwards developed using vibration modes as a basis. The tower stiffness can vary due the reinforcement of the structure. Since it is not easy to foresee their behavior, an uncertainty quantification study appears necessary. Because the nonlinear structures show special sensitivity to dynamic loads, the nominal value of wind velocity (used to construct the wind loads) is also considered as a stochastic parameter. Then, Monte Carlo simulations are performed. The probability distribution functions (PDF) of the results are observed accounting for each parameter separately. Afterwards, both parameters are considered stochastic by means of a joint distribution. Both parameters show very different types of influence on the statistics of the response.
1 INTRODUCTION

For many years, guyed masts have been used to support antennas for radio, TV and other types of communication (Fig. 1(a)). A typical configuration comprises a lattice tower with triangular cross-section (three legs, horizontal and diagonal members). Since the wind load contains energy that interacts with flexible structures, the dynamic response becomes important in the analysis of guyed masts. The mast acts strongly in a non-linear fashion when the guys vary between slack and taut states. The wind load is variable in space and time (during an event) and it is variable from an event to another. Hence, a study considering the wind load as a stochastic event, and the study of the response of the structure to a variety of events with different nominal velocities (that can be considered stochastic as well), is desirable to understand the behavior of the structural system. Also, the column-beam stiffness can be a variable value. For example, in the case of guyed tower, various companies share the use of one structure to install their antennas; this situation requires, in most of the cases, the retrofit of the mast. Some of the strategies consist in the partial reinforcement of the legs of the lattice mast. However, the stiffness of the reinforced mast is not easy to evaluate due to the construction procedures. Thus, a stochastic treatment of the mast stiffness seems appropriate.

![Figure 1](image)

**Figure 1**: (a) Typical guyed tower for mobile signal transmission. (b) Reduced model under study

Despite the large potential of adverse impact, dynamic actions, as wind and earthquakes, are not addressed in detail with exception of special cases (e.g.,[1]). In this work, the stochastic dynamic response of a reduced and simplified model of a guyed tower is analyzed under wind loads. The fluctuating wind component is obtained through the Spectral Representation Method (SRM) [2] starting from a given Power Spectral Density function ($psdf$). The temporal and spatial correlations are taken into account by finding the cross-spectrum and introducing a coherence function. The method yields a temporal record of the fluctuating wind velocity. Combined with the standard recommendations, the fluctuating wind pressure are applied on the mast. Using the classical extended Hamil-
ton’s principle, the equations of motion that govern the transverse vibration are obtained. The nonlinear dynamics of a similar model are studied by Gattulli and Lepidi [3]. Then, a finite element discretization of the linearized equations is used to perform a modal analysis of the structure. Finally, by means of a Galerkin Method, the main nonlinear PDE system is reduced into a 2-DOF model, using two selected normal modes extracted from the previous modal analysis. With this reduced order model (ROM), the nonlinear response of the structure excited by a stochastic wind load is evaluated, considering uncertainties on the stiffness of the mast and the nominal wind velocity.

2 MODEL DESCRIPTION

2.1 Problem formulation

The plane configuration of a guyed column (see Fig 2) is described, referenced to the static equilibrium state, by the cable displacement components $u_c(x_c, t)$ and $v_c(x_c, t)$ along the abscissa $x_c$, and by the beam-column transverse displacements $v_b(x_b, t)$ and $u_b(x_b, t)$ along $x_b$.

![Diagram of Cable-stayed Tower Configurations](image)

Figure 2: Cable-stayed tower configurations: a) static; b) dynamic.

The following assumptions are made: a) both the cable and the beam-column are considered as homogeneous one-dimensional elastic continua obeying a linear stress-strain relationship; b) the equilibrium configuration for the inclined cable is described through a quadratic parabola under the assumptions of small sag to length ratio; c) axial extensions of the cable are described by the Lagrangian strain of the centerline; d) the flexural, torsional and shear stiffness of the cable are negligible; e) the torsional and shear strain of the beam are negligible; f) the nonlinearity of the problem arises from the cable formulation; g) a second order effect due of the axial load (assumed constant) is accounted for in the column equation. The cable formulation follows the approach reported by [3]. Under these assumptions and using the classical extended Hamilton’s principle, the equations of
motion governing the transverse vibration of the guyed column are obtained:

\[
\begin{align*}
V_b : m_b \ddot{v}_b + c_b \dot{v}_b + EI \dddot{v}_b + P_H \dddot{v}_b &= F_{vb}(t, x_b) \\
U_b : m_b \ddot{u}_b + c_b \dot{u}_b + EA_b \ddot{u}_b &= F_{ub}(t, x_b) \\
V_c : m_c \ddot{v}_c + c_c \dot{v}_c - [H \dot{v}_c + EA_c (Y_c' + v'_c) \epsilon_c] &= F_{vc}(t, x_c) \\
U_c : m_c \ddot{u}_c + [EA_c \epsilon_c] &= F_{uc}(t, x_c)
\end{align*}
\] (1)

with the following set of geometric and mechanical boundary conditions:

\[
\begin{align*}
u_c(0) = v_c(l_c) &= 0 \\
v_b(0) = -u_c(0) \cos \theta + v_c(0) \sin \theta \\
u_b(l_b) = v_b(l_b) = u_b(l_b) &= 0 \\
EI \dddot{v}_b(0) &= 0 \\
\ddot{v}_b(0) + (EA_c \epsilon_c + H) \cos \theta + [EA_c \epsilon_c (y_c''(0) + v'_c(0)) + H v'_c(0)] \sin \theta = 0
\end{align*}
\] (2)

Here \( (\cdot)' = d(\cdot)/dx_c \) and \( \dot{(\cdot)} = d(\cdot)/dt \), \( m_b \) and \( m_c \) denote the beam and cable mass per unit length, respectively, \( EI = E_b I_b \) the beam flexural stiffness, \( P_H \) is the component of the initial pretension in the direction of the mast, \( EA_c = E_c A_c \) and \( EA_b = E_b A_b \) the cable and beam axial stiffness, respectively, \( H \) is the mean static tension in the cable, \( Y_c \) is the initial configuration of the cable and, due the hypothesis of small sag to span ratio, \( Y_c(x_c) = 4D(X_c/l_c - (X_c/l_c)^2) \), finally \( \epsilon_c = u_c + y_c' v_c' + 1/2v_c'^2 \) is the expression of the elongation for the cable. Since in cable-stayed structures the beam and the cable materials can generally have different viscous behavior, the model accounts separately for beam and cable transverse damping per unit length, \( C_b \) and \( C_c \) respectively; here an estimate of system damping is assumed to give the conventional values \( c_i = 8.5e^{-3}(2m_i \omega_i) \), with \( i = c, b \) and \( \omega_i \) is a natural frequency of the system. Here \( \omega_i \) is chosen by observation of the modal shapes used in the approximation, i.e. if the modal shape selected is a cable-like mode, then the corresponding frequency is used to the damping term of the cable; the value of \( \omega_i \) were found with \( I_b = 8.33e^{-4}m^4 \).

The values of the constants for the problem are detailed in Tab. 1.

<table>
<thead>
<tr>
<th>( l_b )</th>
<th>( E_b )</th>
<th>( I_b )</th>
<th>( A_b )</th>
<th>( m_b )</th>
<th>( l_c )</th>
<th>( E_c )</th>
<th>( A_c )</th>
<th>( m_c )</th>
<th>( H )</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>2.1e^{11}</td>
<td>4.16e^{-4} - 13e^{-4}</td>
<td>0.04</td>
<td>314</td>
<td>32.31</td>
<td>1.5e^{11}</td>
<td>7.854e^{-5}</td>
<td>0.61</td>
<td>10500</td>
</tr>
<tr>
<td>m</td>
<td>N/m²</td>
<td>m⁴</td>
<td>m²</td>
<td>kg</td>
<td>m</td>
<td>N/m²</td>
<td>m²</td>
<td>kg</td>
<td>N</td>
</tr>
</tbody>
</table>
2.2 Finite Element discretization and Reduced Order Model (ROM).

After stating the weak formulation for the guyed column, the governing system of
equations 1 is discretized by means of an ad-hoc finite element formulation. The column
is modeled using 6-DOF beam elements and the cable using 6-DOF (three node) cable
elements. Figure 3 depicts three selected modal shapes found with a modal analysis.

![Figure 3: Results of FEM modal analysis](image)

Then, by means of a Galerkin Method, the main nonlinear equations system is further
reduced into a 2-DOF model, using two selected normal modes extracted from the previous
modal analysis. With this ROM the nonlinear response of the structure, excited by a
stochastic wind load is evaluated, considering the uncertainty of the stiffness of the mast.
The 2-DOF ROM has been previously verified with an analogous finite element model
from a commercial software and reproduces correctly the amplitude, shape and frequency.

The displacements have been expressed in the modal space as $v = \Phi q$ where the
displacement vector is ordered as $v = \{v_b(x_b, t), u_b(x_b, t), v_c(x_c, t), u_c(x_c, t)\}^T$, the modal
matrix $\Phi = [\phi_1 | \phi_2]$ of the eigenfunctions $\phi_i = \{\phi_{v_b}(x_b), \phi_{u_b}(x_b), \phi_{v_c}(x_c), \phi_{u_c}(x_c)\}^T$ and
the modal amplitude vector is $q = \{q_1(t), q_2(t)\}^T$. Imposing the stationarity of the associ-
ated Hamiltonian, the following nonlinear ordinary differential equations in terms of the
generalized coordinates $q_1$ and $q_2$ are obtained:

$$
\begin{align*}
    m_1 \ddot{q}_1 + a_1 \dot{q}_1 + c_1 q_1 + c_2 q_2 + c_{12} q_1 q_2 + c_{11} q_1^2 + c_{22} q_2^2 + c_{21} q_1 q_2 + c_{122} q_1 q_2^2 + c_{111} q_1^3 + c_{222} q_2^3 &= p_1(t) \\
    m_2 \ddot{q}_2 + a_2 \dot{q}_2 + d_1 q_1 + d_2 q_2 + d_{12} q_1 q_2 + d_{11} q_1^2 + d_{22} q_2^2 + d_{21} q_1 q_2 + d_{122} q_1 q_2^2 + d_{111} q_1^3 + d_{222} q_2^3 &= p_2(t)
\end{align*}
$$

2.3 Wind load

In this work, a wind load with a mean varying with the height ($y$ coordinate, see Fig.
2) in a potential law fashion (see Fig 1(b)) given by the standards, and with a temporal
variation generated by means of the Spectral Representation Method (SRM) [2] is applied on the structure. The wind load takes into account both temporal and spatial correlations. Following the SRM, a signal that verifies the assumptions of a homogeneous, ergodic and Gaussian process with zero mean, and under certain conditions, can be recreated by the following series:

\[ u_y(t) = \sum_{k=1}^{m} \sum_{n=1}^{N} |H_{yk}(w_n)| \sqrt{2\Delta w} \cos[\hat{w}_n t + \theta_{yk}(w_n) + \Phi_{kn}] \]  

where \( \Delta w \) is the frequency interval with which the \textit{psdf} is discretized, \( w_n = \Delta w(n-1) \), \( \hat{w}_n = w_n + \psi_{kn} \Delta w \), \( \psi_{kn} \) is a random value uniformly distributed between 0 and 1, \( N \) is the amount of frequency ranges and, \( \Phi_{kn} \) are the random independent phase angles uniformly distributed between 0 and \( 2\pi \). If the values of \( S_{jk} \) are all real, then the \( \theta_{jk}(w_n) \) are equal to zero. The lower triangular matrix \( H(w) \) is obtained by means of the \textit{Cholesky Decomposition} of the spectral density matrix.

3 UNCERTAINTIES QUANTIFICATION

The uncertainty quantification is performed with two parameters: the bending stiffness of the mast \((EI)\) and the wind nominal velocity \((V_n)\). Monte Carlo simulations are carried out by taking the stiffness of the mast \((EI)\) and the nominal wind velocity \((V_n)\) as stochastic parameters, first separately and then, joined.

The Principle of Maximum Entropy (PME) [4] allows to determine the best PDF that suits with the imposed constraints, and introduces no unwarranted information. The measure of uncertainty of a random variable \( X \) is defined by the following expression

\[ S(f_X) = - \int_D f_X(X) \log(f_X(X)) dX \]  

in which \( f_X \) stands for the PDF of \( X \) and \( D \) is its domain.

Since there is not sufficient statistical information over the stiffness of the mast, here, the response under three assumptions of constraints is studied:

1. If the only constraint are the range values of the random variable, the PME leads to a \textbf{uniform PDF}.

2. If the range values and the mean value are known, the PME leads to a \textbf{truncated exponential PDF}.

3. The application of the PME under the constraints of \textbf{positiveness and bounded second moment} leads to a \textbf{gamma PDF}.

For the \( V_n \) parameter, a \textit{Gumbel} and a \textit{Frechet} distributions are used, since that are commonly used PDFs for extreme winds.
Monte Carlo simulations are performed in order to find the influence of the stochastic parameters, given the selected PDFs, on the response of the structure. To achieve significant statistical results, a convergence study on the standard deviation was first performed to determine the minimum number of realizations of the Monte Carlo simulations. It was found that at least 1000 realizations are required to achieve convergence of the standard deviation.

4 RESULTS

In this work, a guyed beam-column ROM is studied under a stochastic wind load. The total runtime of each experiment is 30 s. In this section, all the results corresponds to the displacements at the top of the mast. The shape of a typical run is depicted in the Fig. 4, despite the main frequency and amplitude of the displacements change for each combination of $EI$ and $V_n$.

![Figure 4](image)

**Figure 4**: Time history of the transverse displacement of the beam-column at point $x_b = 0$ m (tip of the mast) found with the deterministic reduced order model.

4.1 Stiffness of the mast as stochastic variable

In this section, the $EI$ parameter is considered stochastic. A uniform PDF, a truncated exponential PDF and a gamma PDF are respectively used to observe the propagation of uncertainty of this parameter over the local peak displacements ($PD$) statistics. The influence of the $V_n$ is also studied, since the statistics using $EI$ as stochastic parameter are behold for each individual case of $V_n$.

Figure 5 depicts a plan view of the PDF of the $PD$ at the top of the guyed mast. Each vertical strip corresponds to a case of $V_n$. The warmer colors indicate higher probability (and also the location of the modes) while the colder colors indicate the opposite. All the resulting distributions are gamma-like PDFs and the mean and the mode mostly coincide. No differences can be seen when using a uniform PDF or a truncated exponential PDF in the PDF of the results. The gamma PDF instead, shows a smaller dispersion in the results. Nevertheless, the differences are small and it seems that the choice of the PDF for $EI$ does not affect widely the statistics of the peak displacements. The $V_n$ parameter seems to influence the mode and dispersion in a linear fashion (specific figures are not provided for sake of brevity).
4.2 Wind nominal velocity as stochastic variable

In this section, the $V_n$ parameter is considered stochastic. Gumbel and Frechet PDFs
are respectively used to observe the propagation of uncertainty of this parameter over
statistics of the $PD$. The influence of the $EI$ is also studied, since the statistics using $V_n$
as stochastic parameter are considered for each individual case of $EI$.

Figure 6 depicts the PDF of the $PD$ at the top of the guyed mast in a similar fashion as
Fig. 6. Each vertical strip corresponds to a case of $EI$. The differences using the different
PDFs are evident. The PDF of the results using a Gumbel PDF are more disperse and
often multimodalities can be observed (particularly in the third portion corresponding
to lower values of $EI$). On the other hand, using a Frechet distribution for the input
random variable, the PDF of the $PD$ results more compact and Gamma shaped; the
modes and the PDFs are located in a zone of higher $PD$ values, in an agreement with a
Frechet PDF, which have higher probabilities on the zone of higher \( V_n \) while the Gumbel distribution have a more distributed probability among all the values of \( V_n \). Yet, the mode corresponding to the higher value of \( PD \) in the Gumbel case matches the mode of the Frechet case, then, the prediction of the most dangerous displacements is accomplished similarly in both situations. The difference lies in that using a Gumbel PDF yields wider dispersion, since other lower values also has high probabilities of occurrence.

The \( EI \) parameter affects the mean and the dispersion, strongly in the first third region of the values of \( EI \) and then its influence is reduced, but still present.

### 4.3 Wind nominal velocity and stiffness of the mast as joint stochastic variables

Next, both \( EI \) and \( V_n \) are considered stochastic together. The resulting PDFs of \( PD \) are depicted in Fig. 7. Again, it can be seen that the distribution of \( EI \) almost does not affect the PDFs of peak displacements. For the case of Frechet PDF, the result is more compact than the case using a Gumbel PDF. In this study, no multimodalities are observed and the modes do not match, being the more conservative results the ones corresponding to the Frechet PDF.

![Figure 7: PDF of peak displacements using \( EI \) and \( V_n \) as joint stochastic variables. Each subfigure plots both the cases of Frechet and Gumbel and: (a) uniform; (b) truncated exponential; c) gamma.](image)

### 4.4 Probability of exceeding and conditional probability of exceeding a limit value

In this section, the probability that the displacements (\( D \)) (all the displacements of each experiment, not exclusively the peaks) exceed a limit value, arbitrarily fixed on \( D_{lim} = 0.20 \text{m} \), is first evaluated. Since the wind load is a stochastic process, the response of the system (the displacements, in this case) is a stochastic process as well. Then, each run is treated as a stochastic process and the probability of exceeding a limit value, is evaluated for each combination of \( EI \) and \( V_n \) (not considered stochastic in this case).
Since each run is a "temporal" experiment, Fig. 8 can be interpreted as well as the percentage of the total runtime that the structure exceeds the limit value. In Fig. 8 it is noticeable that the variation along a vertical line (fixed value of $V_n$ and different values of $EI$) is lesser when compared to the obtained on a horizontal line (fixed value of $EI$ and different values of $V_n$); it shows that the influence of $EI$ is inferior than $V_n$ on the response, when a specific value (or range) is studied.

![Figure 8: Probability that the displacements exceed a limit value of 0.20 m, for each pair of values of $EI$ and $V_n$.](image)

Next, the conditional probability of exceedance a limit value of $D_{lim} = 0.20$ m, now considering $EI$ and $V_n$ stochastic, is studied. In this instance, the probability of exceeding a given limit value in some parameter combination is assessed, and furthermore, if that parameter combination is probable. The results are plotted in Fig. 9, for all the combinations of PDF studied in this work. As before, the warmer colors indicate a higher concentration of probabilities and the colder, the opposite. As can be seen, when the probability of the parameters are involved, the results vary widely from when they are considered deterministic (Fig. 8).

The results have different shapes when comparing the cases of PDF used for $EI$. When the uniform PDF or the truncated exponential PDF are used, the highest probabilities are concentrated in the zone of lesser values of $EI$ (more dispersion is observed using the uniform PDF, as expected). Nevertheless, the results with both cases of PDF are very disperse. When the gamma PDF is chosen, the probabilities are focused in the zone of reference $V_n$ (55 m/s) and the mean value of $EI$.

The use of a Gumbel PDF (plots (a), (b), (c) in Fig. 9) is characterized for the presence of a larger dispersion in the direction of the $V_n$ axis, when compared with the results using a Frechet PDF (plots (d), (e), (f) in Fig. 9).

This study reveals that, for this system, the lesser dispersion in the input PDFs, the lesser dispersion is observed in the conditional probabilities of exceedance. The fact that
the different choices for the parameter PDF gives qualitatively different results, indicates that a better knowing of the statistics of the parameters leads to a better characterization of the response.

5 FINAL REMARKS

In this work, a nonlinear formulation of a cable-stayed mast structure is presented and the model is used to perform a stochastic dynamic analysis. First, a deterministic model is stated and the governing system solved via a finite elements discretization. A 2-DOF reduced order model (ROM) of the system is then constructed using the eigendata from the previous model. Once calibrated, the ROM was used to perform an uncertainty propagation study using the stiffness of the mast and the nominal wind velocity as stochastic parameters and a stochastic load to simulate the wind actuating on the mast. The loads take into account spatial and temporal correlations. The PDFs of $EI$ were found by means of the Principle of Maximum Entropy under various assumptions of constraints, given the available statistical data. For the $V_n$ parameter two commonly used distributions were selected. Monte Carlo simulations were employed to find the realizations.

The results were studied statistically. First, the parameters were considered as random variables separately. When $EI$ is considered stochastic, it is found that the PDF of the local peak displacements present similar gamma-like shapes for all the three cases of PDF
studied, suggesting that the statistic of $EI$ have a small importance in the statistic analysis of the response; the $V_n$ parameter affects the mode and the dispersion of the results in a linear fashion. When $V_n$ is considered stochastic, multimodalities and a larger dispersion are found when using a Gumbel PDF, while when using a Frechet PDF the results are more compact but also conservative; the influence of $EI$ in the shape and dispersion of the resulting PDFs is remarkable in the first third portion of values of $EI$. When both $EI$ and $V_n$ are considered stochastic, and the PDF of the peak displacements is studied, similar conclusions as before can be stated, but no multimodalities are observed. The influence of the PDF used for $EI$ becomes noticeable when the probability for the displacements of exceeding a limit value is studied. Further studies in order to statistically characterize the $EI$ are probed necessary.

6 ACKNOWLEDGEMENTS

The authors acknowledge the financial support from CONICET, MINCyT and UNS (Argentina); CAPES, CNPq, and FAPERJ (Brazil).

REFERENCES


PRELIMINARY EVALUATION WITH ENSEMBLE PREDICTION FOR THE SUPIM MODEL

Fernando Emilio Puntel¹, Adriano Petry¹, Haroldo Fraga de Campos Velho² and Jonas Rodrigues de Souza³

¹ Southern Regional Space Research Center
National Institute for Space Research, Santa Maria, RS, Brazil
Contact: [fernando.puntel, adriano.petry]@inpe.br, phone (+55 55) 33012011
Av Roraima, s/n, campus UFSM, prédio do INPE, sala 2000
P.O. Box 5021, CEP 97105-970, Santa Maria (RS), Brazil

² Associated Laboratory for Computing and Applied Mathematics
National Institute for Space Research, São José dos Campos (SP), Brazil
Contact: haroldo@lac.inpe.br, phone (+55 12) 3945-6547

³ Aeronomy Division
National Institute for Space Research, São José dos Campos (SP), Brazil
Contact: jonas@dae.inpe.br, phone (+55 12) 3945-7162

Key words: VTEC, SUPIM, ensemble forecasting.

Abstract. An operational version from the Sheffield University Plasmasphere-Ionosphere Model (SUPIM) was developed to the Brazilian Space Weather program in INPE - Brazil (http://www2.inpe.br/climaespacial/). The SUPIM is a physics-based model computer code describing the distribution of ionization. The model applies electric-magnetics coordinates, where a conformal map is employed to transform the geographic (Cartesian) coordinates into electric-magnetics ones. On this transformed coordinate, the computation is performed in a 2-dimensional plane aligned with Earth magnetic field lines, with fixed magnetic longitude coordinate. The SUPIM is executed at different longitudes. Several adaptations were developed to the operational version to run daily, including: parallel version for the code (better CPU-time performance), adding an artificial exponential decay for ion–neutral collisional frequency for high altitudes (better code stability), a module for data assimilation (by Newtonian relaxation method). The operational system runs every day, forecasting almost 24 h ahead for the VTEC (vertical total electronic content) values on the South America. Preliminary tests by using an ensemble prediction scheme are under investigation. Ensemble forecasting belongs to the class of Monte Carlo approach, where multiple numerical predictions are carried out. There are several techniques to calculate the ensemble members. Our approach adopts one strategy to employ slightly different initial conditions. The ensemble forecasting allows to identify zones of...
low and high predictability, as well as to compute a confidence interval. The latter issue is a quantification of uncertainty associate to the prediction.

1 INTRODUCTION

The prediction using mathematical models by using numerical methods is a remarkable scientific achievement. Such framework has been developed initially for weather forecasting the mid-twentieth century (see [14]), and became a methodology adopted to forecast ocean circulation [9], air pollution [7], electronic content of the atmosphere [15].

Space weather embraces solar physics coupled with Earth system, mainly magnetosphere and ionosphere dynamics. The INPE (National Institute for Space Research, Brazil) has adapted the SUPIM (Sheffield University Plasmasphere-Ionosphere Model) computer code for producing maps of electronic content in the atmosphere [15] 24 hours ahead (see also: http://www2.inpe.br/climaespacial/portal/tec-supim-previsao/). The latter activity is part of the Brazilian Space Weather Program, officially starting since 2007.

However, our models are only an approximation of reality. Other errors are also present: the initial condition is not described with precision, some stochastic forcings are not computed and/or known, the numerical method contains errors. In addition, Lorenz has shown that some dynamical systems can be extremely sensitive to initial conditions (these are called chaotic systems [13]). There is a necessity to previously evaluate the prediction reliability. This study is called predictability, in other words, we want to establish or quantify the degree of uncertainty of the numerical prediction. One strategy to quantify such uncertainty is to employ the ensemble prediction.

Ensemble prediction is a Monte Carlo technique.

Several methods have been applied to compute the ensemble. One of them deals with different initial conditions. This is the approach used here. The SUPIM is executed under different initial conditions, allowing to calculate some statistical properties as ensemble mean and variance. The variance can be understood as the uncertainty measure. Few words about ensemble prediction are expressed in the next section. Section 3 presents a brief description on the SUPIM. The methodology employed and the numerical experiments carried out are in Section 4. The final section addressed some final remarks and conclusion.

2 ENSEMBLE METHODS

The ensemble members can be obtained using different strategies: with a set of different initial conditions [8], different numerical values for model parameters, or by employing different models [5, 11, 12].

The method named multimodel-superensemble [11] is an ensemble-based approach, where the response from several models is combined, and a weighting average is cal-
culated. Cane and Milelli [4] applied the latter method to predict the precipitation on the Piemonte (Italy) region. The precipitation forecasting was improved. Similar approach was used by Santos et al. [16] for predicting the precipitation field on the Amazon region, with better response with BRAMS model.

The ensemble forecasting more usually employed by the operational prediction centers is a form of Monte Carlo analysis by dealing with slightly different initial conditions, and the numerical prediction is performed considering all initial condition[8]. Therefore, a set of predictions is obtained, where an individual prediction is called member of the ensemble.

The dynamical system can drop on different scenarios:

(i) unstable dynamical system, where even very small difference on initial conditions can produced strong divergence between two orbits.

(ii) stable dynamical system, where very different initial conditions converge to a small region in the phase space.

As already mentioned, the ensemble technique allows to evaluate the system response that deals with imperfect model, observational errors, some forcing terms badly treated or even neglected, for quantitatively measuring the reliability of the forecast. Thus, the ensemble process gives more robust prediction and more fault-tolerant [3].

3 THE SUPIM PREDICTION SYSTEM

The SUPIM-DAVS is a forecasting system of ionospheric dynamics based on SUPIM model. The SUPIM is a model from the first principles for the Earth ionosphere and plasmasphere. This model has been developed in the last three decades [1, 2, 17].

In SUPIM code, the equations of continuity, momentum, and energy are time dependent. To calculate the values for density, temperature, and flow of electrons and ions O+, H+, He++, N2+, O2+, NO+, all equations are solved along the magnetic field. The model also encodes several physical and chemical processes [15].

The SUPIM-DAVS is operationally executed since March 2011 in a computer system (clusters) from the South Regional Center for Space Research (CRS/INPE), Santa Maria (RS), Brazil. Figure 1 shows a flow-chart of the SUPIM-DAVS execution.

The first procedure performed by the SUPIM-DAVS system is obtaining solar flux data. Eletroctronic content observations, used for data assimilation, can be obtained from different sources like ionograms from the National Oceanic and Atmospheric Administration (NOAA) web-page, INPE Space Weather webservice, or from the local server. The next step is to validate the solar flux data through SOLAR2000 code.

Using the obtained solar flux data, and setting up parameters that define the model for Earth’s magnetic field, thermospheric winds, and the simulation date, the SUPIM is executed. The DAVS is divided into 3 parts: data assimilation, temporal adjustment on
simulation results based on longitudinal offset, and post-processing (including homogeneous grid interpolation). Finally, frames of images are generated using GrADS (Grid Analysis and Display System) software.

The SUPIM is prepared to use prior forecast data. Such data register 5 different properties: electron concentration and its temperature, ionic concentration and its temperature, and thermospheric wind. The latter 3 properties are indexed to each of the 7 simulated ions. If there is no previous data of electronic concentration, the system performs an iterative process to determine an equilibrium condition. After reaching convergence for the latter iteration, the system is ready to perform the simulation.

4 EXPERIMENTS AND RESULTS

All experiments were carried out for the day 12/Dec/2014. Two types of experiments were performed. Firstly, looking at the average of the VTEC (Vertical Total Electron Content) over the whole South America map. Another one, VTEC values at 6 points geographically spaced over South America map, as shown in Figure 2. For type-2 experiment, 6 sampled points were selected to evaluate local behavior, just because the global average could hide a local anomalous behavior.
4.1 Methodology

For implementing the ensemble method, it was necessary to determine different initial conditions. To generate the initial conditions, the properties calculated by SUPIM were changed by adding a random noise.

The noise were added on only two properties: electron content and ionic concentration for each one of 7 ions. The other properties were not changed, because their variation has no no significant impact on TEC values.

Five different simulations were performed to implement the method:

(a) Simulation without noise.
(b) Simulation with positive maximum noise level of 30%.
(c) Simulation with negative maximum noise level of 30%.
(d) Simulation with positive maximum noise level of 50%.
(e) Simulation with negative maximum noise level of 50%.

4.2 Results

Figure 3 illustrates the simulations for the time evolution of the TEC average over South America. The red lines represent noisy simulations and black points the simulation without noise. The latter points also show the standard deviation interval (vertical lines over the points). The interval is calculated with the noisy values. This gives us a notion of dispersion among different simulations. It is clear the convergence of curves for all 5 simulations. For this simulation, the snapshot indicates a stability condition.

For the experiment using 6 points, shown in Figure ??, the evolution of TEC values also showed similar behavior to the previous experiment, indicating local stability for the marked points. However, the convergence rate varied for each location. In location (b), convergence was slower than in other locations evaluated. Location (c) achieved small values for standard deviation after few hours of simulation. These differences in convergence evolution can be attributed to the physical process modeled, where regions of very high electronic concentration follow the curvature of the Earth’s magnetic field in the region.

5 CONCLUSIONS

The simulation for ensemble members converged for the same value. In other words, our experiments have showed a prediction with high reliability (low uncertainty), which is measured by the decrease of the variance along time.

The performed experiments are not to be considered as a representative ones. The goal is to show, by worked examples, the methodology under development applied to quantify the uncertainty linked to the prediction, indicating the degree of reliability of the forecast.
For all experiments, the SUPIM-DAVS performed stable simulations, even with the presence of noise in the data, generating acceptable results.
Figure 4: Ensemble TEC evolution for 6 points on South America.
ACKNOWLEDGEMENT

The authors wish to thank CNPq and FAPESP, Brazilian agencies for research support.

REFERENCES


THERMO-MECHANICAL MULTI-OBJECTIVE BIDIRECTIONAL EVOLUTIONARY STRUCTURAL OPTIMIZATION USING WEIGHTED SUM METHOD FOR MEAN COMPLIANCE AND HEAT CONDUCTION PROBLEM

Sergio Quispe Rodríguez\textsuperscript{1}, Renato Pavanelo\textsuperscript{2}

\textsuperscript{1} Universidade Estadual de Campinas, Departamento de Mecânica Computacional
Campinas, S.P. Brazil
sergioquispe@fem.unicamp.br

\textsuperscript{2} Universidade Estadual de Campinas, Departamento de Mecânica Computacional
Campinas, S.P. Brazil
pava@fem.unicamp.br

**Key words:** Structural Topology Optimization, Thermo-Mechanical Structures, Design-dependent loads, Multi-objective Optimization, Weighted sum Method.

**Abstract.** This work proposes an approach of the Bidirectional Evolutionary Structural Optimization (BESO) method to deal with mean compliance minimization and heat conduction optimization in continuum linear static structures subjected to thermo-mechanical design-dependent loads considering a uniform temperature variation for each element as a result of steady heat conduction problem. Based on the weakly coupled thermo-elastic finite element theory, the equivalent nodal thermal load is obtained and applied to the elastic field as a physical body force in the mean compliance minimization problem. The thermo-mechanical analysis is accounted using the sequence coupling method, and a sensitivity analysis using the adjoint method for mean compliance minimization. For the multiobjective thermomechanical optimization, the weighted sum method is applied. The present work analyses the influence of the weighted sum method parameters in the thermomechanical coupling optimization and in the optimal topologies obtained using BESO method. Numerical examples are presented to show the feasibility of the present approach. Good agreement has been observed when compared with results found in the literature.

1 INTRODUCTION

Structural optimization, in particular the layout optimization, has been identified as the most challenging and economically the most rewarding task in structural design [1].
The task is to find an optimal structure with the best performance under some restrictions. Up to now, most articles in structural topology optimization have focused their studies on the elastic field or the thermal field only, as independent ([2, 3, 4, 5]). There are few works that analyse the coupled thermo-elastic effect ([6, 7]).

The concentrations of high temperatures in a structure affects its performance and resistance because of the thermal stresses that implies, without mentioning the modification of the material properties involved in this temperature conditions. For this reason, structures that overcome this problems the best way as possible have great importance and have been getting a great attendance in latests works. This works aims to depict the structural optimization, the topology optimization in particular, using the Bidirectional Evolutionary Structural Optimization (BESO) method to find more rigid structures with a temperature distribution as uniform as possible, obtaining by a consequence a cooler and more rigid structure for an available quantity of material, considering an environment with mechanical and thermal restrictions. Based on the BESO procedure presented in [8], which minimize compliance in structures submitted to thermo-mechanical loads, the heat conduction is also studied and applied as a objective function as well, proposing a multi-objective BESO procedure for this applications.

2 Thermo-Mechanical Analysis

In the thermo-mechanical analysis, the thermal loads \((ft)\) due to thermal expansion that continuum structures generate in an uniform temperature field, and heat distortion due to that thermal loads [9], as shown in Fig.(1), are studied. It is shown that the thermal load due to thermal expansion is a body load, this concept mainly refers to the basic feature “no material, no body force” [10], physically, the location and direction of the body force remain often unchanged whereas the magnitude varies with the material distribution. This definition also suits for the design-dependent loads. The thermal nodal loads in a single element (part of the whole mesh) are shown in Fig.(1). The thermal nodal load vector can be expressed by [11]:

\[
\{fh_i\} = \int\int_{\Omega} [B]^T[D] \alpha_T \Delta T [1 \ 1 \ 0]^T h dx dy
\]

where \(\{fh_i\}\) is the nodal load vector applied in the \(i\)th element, \(\alpha_T\) is the linear thermal dilatation coefficient and \(\Delta T\) is the temperature variation in the element. \([B]\) is the strain matrix and \([D]\), the elasticity matrix. Those matrices are expressed by Eq. (2). \(E\) is the material Young’s modulus, \(\mu\) is the Poisson coefficient and \(N_i\) are the shape functions of quadrilateral elements.

\[
[B] = \begin{bmatrix}
\frac{\partial N_i}{\partial x} & 0 \\
0 & \frac{\partial N_i}{\partial y}
\end{bmatrix}, \quad [D] = \frac{E}{1-\mu^2} \begin{bmatrix}
1 & \mu & 0 \\
\mu & 1 & 0 \\
0 & 0 & (1-\mu)/2
\end{bmatrix}
\]
In this work, a modified BESO procedure, presented in [8], is used. This procedure considers thermal design dependent body loads in the elastic field, for an uniform temperature variation distribution. The sequential coupled method is used to solve the thermo-mechanical effect in the mean compliance minimization problem. First, the temperature distribution in a steady state is obtained considering the thermal effects. After that, the dilatation load vector is calculated as nodal body forces applied in the elastic field. Finally, using the finite element method, the nodal displacements caused by the effects of the thermal and elastic fields (thermal and concentrated loads) is analysed. Therefore, the global finite element equation is:

$$\begin{align*}
[K] \{u\} &= \{f\} = \{fm\} + \{fh\} \\
\text{where } [K] \text{ is the global stiffness matrix, } \{u\} \text{ is the global displacement vector, } \{fm\} \text{ is the global concentrated load vector, } \{fh\} \text{ is the global thermal load vector, caused by the temperature variation (\(\Delta T\)). From Eq.}(3), \text{ the global displacement vector can be obtained, then the mean topology optimization problem for minimizing the mean compliance, would be expressed by:}
\end{align*}$$

Find: \( \{x\} = \{x_1, x_2, ..., x_i\}^T \in \mathbb{R}^n \)

Minimize: \( \displaystyle C = \frac{1}{2} \{f\}^T \{u\} \)

Subject to: \( \begin{align*}
[K] \{u\} &= \{fm\} + \{fh\} = \sum_{i=1}^{n} \{fm_i\} + \sum_{i=1}^{n} \{fh_i\} \\
\int_{\Omega} x d\Omega &\leq FV_0 \\
0 < x_{\min} &\leq x_i \leq 1
\end{align*} \)

where \( \{x\} \) is the vector of the design variables, \( \{x_i\} \) is the elemental design variable, \( x_{\min} \) is the minimum possible value of the design variable, \( n \) is the number of elements in the finite element (FE) mesh, \( V_0 \) is the initial volume (total design domain), and \( F \) is the final volume fraction (used as a given restriction in the optimization problem).
3.1 Sensitivity Analysis

The basic formulation of the topology optimization problem consists on the minimization of the mean compliance of the body thermal forces and concentrated forces, subjected to a given volume fraction, Eq.(4). For the sensitivity analysis, the direct and adjoint methods [12] can be used. The direct method was chosen, differentiating the mean compliance:

\[
\frac{dC}{dx_i} = \frac{1}{2} \{f\}^T \frac{d\{u\}}{dx_i} + \frac{1}{2} \frac{d\{f\}^T}{dx_i} \{u\}
\]

(5)

It should be pointed out that the sensitivity of the mean compliance experiences a non-monotonous behaviour, meaning, it can be either positive or negative. Therefore, it could change sign with the value of the design variable.

3.2 Material interpolation model and sensitivity number

The analysis of the topology optimization including density-dependent body loads, is not a direct extension of the classical design problem. There are some particularities and difficulties in the solution of the related compliance minimization problem and especially in the sensitivity analysis. The most relevant difficulties presented are the parasitic effect for low densities and the non-monotonous behaviour of the objective function [13]. To overcome these difficulties, the RAMP (Rational Approximation of Material Properties) interpolation scheme is used [14]. Then, the interpolation function \( \phi(x_i) \), the Young modulus and the thermal dilatation coefficient of the material model design, become,

\[
\phi(x_i) = \frac{x_i}{1+q(1-x_i)}; \quad E_i = \phi(x_i)E^0; \quad \alpha_T = \phi(x_i)\alpha^0_T
\]

(6)

respectively. where \( E^0 \) denote the Young modulus of the solid material. \( q \) is the penalty factor which is larger than 0 for the topology optimization problem. Substituting Eq.(1) and Eq.(6) into Eq.(4), we obtain:

\[
\{fh\} = \sum_{i=1}^{n} \phi(x_i)^2 \{fh^0_i\}; \quad \{fh^0_i\} = \iint_{\Omega} [B]^T[D^0]_{\alpha_T} \Delta T \begin{bmatrix} 1 & 1 & 0 \end{bmatrix}^T h dxdy
\]

(7)

where \( \{fh^0_i\} \) is the thermal nodal load vector for solid elements. Returning to Eq.(5), the derivative of the mean compliance can be expressed by [8]:

\[
\alpha_i = \frac{-(1+q)}{2} \left( \frac{dC}{dx_i} \right)
\]

\[
\alpha_i = \begin{cases} \{u_i\}^T[K_i]\{u_i\} - 4\{u_i\}^T\{fh^0_i\} \quad &x_i = 1 \\ \left( \frac{x_i}{1+q(1-x_{min})} \right)^4 \frac{1}{1+q(1-x_{min})} \frac{1}{(1-q(1-x_{min}))^2} \left[ \frac{1}{1+q(1-x_{min})} \right] \end{cases} 
\]

(8)

\[ x_i = x_{\text{min}} \]
It can be pointed out that the elemental sensitivity of the mean compliance depends on the value of the penalty factor \((q)\), when a high penalty factor is used, the Young’s modulus of soft elements become insignificant and the interpolation scheme approximates to a 0/1 approach, then the variable \(x_i\) becomes discrete where the soft elements (elements with design variable less than 1) are considered to be totally eliminated from the structure. That is the definition of the hard-kill BESO problem. Then, making \(q \to \infty\), and substituting in Eq.(8), the expression that approximates the sensibilities for hard-kill problem results:

\[
\alpha_i = \begin{cases} 
  \{u_i\}^T[K_i]\{u_i\} - 4\{u_i\}^T\{fh^0_i\}; & x_i = 1 \\
  0 & x_i = 0 
\end{cases} 
\]  

where \(x_{\min} = 0\). The above equation implies that the hard-kill BESO method for structures with design-dependent thermo-mechanical loads has the elemental sensitivity number as expressed in Eq.(9).

4 BESO method for heat conduction optimization problem

In this section, the topology optimization for the heat conduction problem is analysed, where the objective function to be minimized is the heat potential capacity \((C_T)\). Here the task of the topology optimization is to find out such an optimal configuration of the structure that produces the least heat for a prescribed volume of available solid materials. In other words, the temperature distribution will be minimized globally [15]. The ESO method method was studied in [16, 5] for the heat conduction topology optimization problem, and proposing an extended procedure that also reduces the temperature in specific regions of the structure. More recently, a BESO procedure of heat conduction topology optimization dealing with design-dependent heat loads was proposed in [15]. In this work, the BESO method proposed is based in this last one. For the heat conduction, the finite element equilibrium equation corresponds to

\[
\{P\} = [k]\{T\} 
\]

where \(\{P\}\) is the thermal load vector, \(\{T\}\) is the nodal temperatures vector and \([k]\) is the symmetric thermal conductivity matrix. Analogously to the mean compliance minimization problem, the design variable is defined as a binary variable which indicates the presence (1) or absence (0) of an element. Here, the heat capacity \((C_T)\) is the objective function, the optimization problem can be defined by

Find: \(\{x\} = \{x_1, x_2, ..., x_i\}^T \in \mathbb{R}^n\)

Minimize: \(C_T = \frac{1}{2}\{P\}^T\{T\}\)

\(\{P\} = [k]\{T\}\)

Subject to: \(V^* - \sum_{i=1}^{N} v_i x_i = 0\)

\(x_i = 0 \vee x_i = 1\)

(11)
where $x_i$ is the binary design variable, $v_i$ is the element volume and $V^*$ is the final structure volume.

### 4.1 Sensitivity Analysis

In this BESO procedure, the heat loads are not design dependent and the design variable used is a discrete binary variable, then the penalize factor $p$ from the SIMP (Solid Isotropic Material Penalization) interpolation scheme has no effect. Then, a linear interpolation of material properties is used, where:

$$ [k_i] = x_i[k_0^i]; \quad x_i = \begin{cases} 1; & \text{solid element} \\ 0; & \text{void element} \end{cases} $$

where $[k_i]$ is the elemental thermal conductivity matrix and $[k_0^i]$ is the elemental thermal conductivity matrix for solid elements. Then, the expression for sensibility numbers becomes

$$ \alpha_i = \{T_i\}[k_i^0]\{T_i\}; \quad \alpha_i = \frac{1}{x_i\{T_i\}[k_i]\{T_i\}} $$

### 5 Multi-objective optimization using BESO method

There are many real applications in structural optimization that demand the designer to consider two or more objective functions at the same time. These problems are called or denominated as “Multi-objective” optimization problems or vectorial optimization problems and can be defined by

$$ \begin{align*} 
\text{Min:} & \quad [F(\{x\})] = [F_1(\{x\}), F_2(\{x\}), \ldots, F_k(\{x\})] \nonumber^T \\
\text{Subject to:} & \quad g_j \leq 0, j = 1, 2, \ldots, m; \\
& \quad h_i = 0, l = 1, 2, \ldots, e
\end{align*} $$

where $k$ is the number of objectives functions, $m$ is the number of inequality restrictions and $e$ is the number of equality restrictions. $\{x\}$ is the vector of the design variables.

There are several types of solution methods in the literature, in this work we use the “Weighted Sum Method” which is a scalarized multi-objective solution method, which combines all the objective functions in only one.

#### 5.1 Weighted Sum Method

The most common approach to solve multi-objective optimization problems is the so called Weighted Sum Method, this approach is defined by

$$ U = \sum_{i=1}^{k} w_i F_i(x) $$
where $U$ is the resultant objective function from the multi-objective problem and $\{w\}$ is the weights vector, which is frequently defined by the designer in order to represent the preferences or priorities of each single objective function [17]. Usually, the sum of these weights is the unity ($\sum_{i=1}^{k} w_i = 1$) and $w_i > 0$ when the objective functions are normalized.

5.2 **Bi-objective thermo-mechanical topology optimization**

The task of this type of topology optimization is to get stiffest structures with a temperature distribution as uniform as possible. Using the weighted sum method, the mean compliance and the heat potential capacity are the objective functions to be optimized in this problem. The declaration of the bi-objective thermo-mechanical topology optimization is expressed by

$$\text{Minimize:} \quad C = \frac{1}{2}\{f\}^T\{u\} \land C_T = \frac{1}{2}\{P\}^T\{T\}$$

$$[K]\{u\} = \{fm\} + \{fh\} = \sum_{i=1}^{n}\{fm_i\} + \sum_{i=1}^{n}\{fh_i\}$$

Subject to:

$$[k]\{T\} = \{P\}$$

$$V^* - \sum_{i=1}^{N} v_i x_i = 0$$

$$x_i = 0 \lor x_i = 1$$

In order to minimize these two objective functions, a new objective function is presented ($C_{TM}$) which combines the mean compliance and the heat potential capacity, this new objective function can be expressed as

$$C_{TM} = w_1 C + w_2 C_T$$

where $C_{TM}$ is the objective function for the thermo-mechanical bi-objective topology optimization. In Eq.(17), the objective functions are not normalized, the normalization proposed in this work is carried out in the sensibility analysis.

5.2.1 **Sensitivity Analysis**

The sensitivity analysis in bi-objective thermo-mechanical topology optimization is made taking in consideration the sensitivity expressions for the single objective function problems presented before. From Eq.(9) and Eq.(13), the sensitivity expression for this case is obtained by

$$\alpha_i = \begin{cases} 
  w_1 \left( \frac{1}{2} \{u_i\}^T[K_i]\{u_i\} - 2\{u_i\}^T\{fh_i^0\} \right) + w_2\{T_i\}[k_i]\{T_i\}; & x_i = 1 \\
  0; & x_i = 0
\end{cases}$$

Then, making the normalization in the sensitivity numbers for each objective functions,
\[ \alpha_1 = \begin{cases} \{u_i\}^T[K_i]\{u_i\} - 4\{u_i\}^T\{fh_i\} & x_i = 1 \\ 0 & x_i = 0 \end{cases} \]

\[ \alpha_2 = \begin{cases} \{T_i\}[k_i]\{T_i\} & x_i = 1 \\ 0 & x_i = 0 \end{cases} \]

Logo: \[ \{\alpha\} = w_1 \left( \frac{\{\alpha_1\}}{\max(\{\alpha_1\})} \right) + w_2 \left( \frac{\{\alpha_2\}}{\max(\{\alpha_2\})} \right) \]

where \( \{\alpha_1\} \) is the vector that contains the sensitivities of the mean compliance and \( \{\alpha_2\} \) the vector that contains the sensitivities of the heat capacity. \( w_1 \) and \( w_2 \) are the weights for each respective objective function. Then the resultant elemental sensitivity vector denominated \( \{\alpha\} \) as shown in the Eq.(19).

5.3 Example of a short cantilever submitted to thermal restrictions

The effect of the chosen weights on the resultant topologies is analysed presenting a typical topology optimization problem. The short cantilever problem is a classical example used in the literature for validation of programs and for results comparison, the solution obtained by BESO optimization for stiffness maximization and the one used to validate the implemented algorithm is shown in Fig.(5.3a). In this paper, certain temperature restrictions and loads are also considered. The task of the optimization problem is to minimize the mean compliance and the heat capacity using several sets of weights values in order to compare the resultant topologies.

A rectangular plate of 80mmx50mmx1mm is considered, clamped at its left side and pulled by a concentrated force in the middle of the right side, as seen in Fig.(2). An imposed temperature variation of 0\(^\circ\)C and a punctual heat generation of \( q_G = 0.65 \text{W/mm}^3 \) in the middle right side is also considered. The following material properties are assumed: Young’s modulus \( E=210\text{GPa} \), Poisson’s ratio \( \mu=0.3 \) and thermal linear dilatation coefficient \( \alpha = 3 \dot{e} - 06/\circ\text{C} \). The BESo method parameters considered are shown in the Table (1).

<table>
<thead>
<tr>
<th>Variável</th>
<th>Descrição</th>
<th>Valor</th>
</tr>
</thead>
<tbody>
<tr>
<td>( V^* )</td>
<td>Fração de volume final</td>
<td>0,5</td>
</tr>
<tr>
<td>( ER )</td>
<td>Razão de evolução</td>
<td>1%</td>
</tr>
<tr>
<td>( AR_{\text{máx}} )</td>
<td>Máxima razão de admissão</td>
<td>5%</td>
</tr>
<tr>
<td>( r_{\text{min}} )</td>
<td>Raio do filtro</td>
<td>3mm</td>
</tr>
<tr>
<td>( \tau )</td>
<td>Tolerância da convergência</td>
<td>0,1%</td>
</tr>
<tr>
<td>( N )</td>
<td>Parâmetro da convergência</td>
<td>4</td>
</tr>
</tbody>
</table>

The design domain and restrictions are illustrated in Fig.(2). First, the heat conduction topology optimization problem is first solved, considering only the heat capacity as the
objective function, the result of the optimization process using BESO method as well as the temperatures profiles, in the initial domain and in the resultant structure, are shown in Fig.(3). It is important to note that even when the final structure presents a higher maximum temperature than the initial domain, the distribution of these temperatures is as more uniform as possible, meaning that any other structure with the same quantity of available material would not present a more uniform temperature distribution.

![Figure 2: Design domain and restrictions: a) In the elastic field; b) In the thermal field.](image)

![Figure 3: Temperature profiles for the initial and optimized structures for the heat conduction topology optimization problem shown in Fig.(2b).](image)

The multi-objective problem takes place in Fig.(5.3). Different sets of weight values were used in order to get different topologies. It is obvious, from Fig.(5.3), that the maximum temperature of the structure decreases when the value of \( w_2 \) increases, it means that a higher value of the weight involving the heat capacity gives it more preference in the bi-objective optimization process and in the criterion of elements addition/removal process. To get a more adequate comparison, the mean compliance of these structures is analysed, obtaining values of \( 7.7710e05 \text{Nmm}, 7.7829e05 \text{Nmm}, 7.7842e05 \text{Nmm}, 7.7947e05 \text{Nmm}, 7.9023e05 \text{Nmm} \) and \( 2.4839e06 \text{Nmm} \) in the order that they are presented in Fig.(5.3). Coming to the conclusion that when the value of the priority of the heat capacity (\( w_2 \)) increases, the minimization of mean compliance is getting less attention in the addition/removal process.
Sergio Quispe Rodríguez, Renato Pavanello

(a) \( w_1 = 1 \) e \( w_2 = 0 \)

(b) \( w_1 = 0.8 \) e \( w_2 = 0.2 \)

(c) \( w_1 = 0.6 \) e \( w_2 = 0.4 \)

(d) \( w_1 = 0.4 \) e \( w_2 = 0.6 \)

(e) \( w_1 = 0.2 \) e \( w_2 = 0.8 \)

(f) \( w_1 = 0 \) e \( w_2 = 1 \)

Figure 4: Temperature profiles for the optimized structures considering different set of weight values. The same temperatures scale is used in order to get a more clear view of the maximum temperature zones in each structure and a better comparison too.

To observe this comparison between the optimized structures, and the effect that the value of weights has in them, more clearly. Fig.(5.3) is presented, where the nodal resultant displacement is plotted in the same scale to compare the deformation in the structures from Fig.(5.3a) and Fig.(5.3f), it is obvious that the structure where the mean compliance is the only objective function does not deform in the same proportion as the structure which has the heat potential capacity as the only objective function, this occurs because of the chosen priorities of the objectives functions, as explained before.

6 CONCLUSIONS

- It is important to mention that all the computations were performed in a simple PC with an Intel Core i7-4770 CPU processor and 16.0GB of RAM.
The BESO procedure proposed in this work aims to minimize the heat potential capacity and the mean compliance. Being impossible to obtain a single minimum point of those two objective functions at the same time, the weighted sum method is studied, giving a priority of each objective function. This priority of the objective functions is frequently chosen by the designer and is a particular value for every specific problem. The value of this priorities of weights chosen, implies that one objective function is going to be “beneficed” in the removal/addition of elements process. It means that this objective function would be minimized in a bigger proportion than the other one. This was observed and demonstrated with a practical example.

The structural optimization for thermo-mechanical structures, as presented in this paper, has a very strong interdependency in the choice of the optimization parameters and design variables (e.g. the design domain, concentrated forces, temperature variation), meaning that when a optimization parameter or design variable is modified, the final topology would be different, then the final topology obtained highly depends of the problem analysed.

The thermo-mechanical optimization considering the mean compliance as an objective function has demonstrated a big versatility and adequate convergence for problems with fixed mechanical loads numerical and proportionally bigger than the thermal loads from the temperature variations, but when the thermal loads are considerably bigger than mechanical loads, this criterion becomes not accurate and convergence problems arrives. Those problems should be studied in future works and other criterion would be used in order to compare the obtained results.

The numerical results show that the proposed BESO method is efficient and convergent and solid-void solutions can be achieved for a variety of thermal distributions in the optimization problems. This modified BESO procedure also presents an advantage from the SIMP method in the clarity of the solid-void interface.
ACKNOWLEDGEMENTS

This work was supported by the National Council for Scientific and Technological Development (CNPq) [grant 132742/2013-6] and the Sao Paulo Research Foundation (FAPESP) [grant 2013/08293-7].

REFERENCES


INTEGRATING EVOLUTIONARY OPTIMIZATION, STATISTICAL CLASSIFICATION AND ANALYSIS OF RELIABILITY, COSTS AND TECHNICAL LOSSES INTO A COMPUTATIONAL SYSTEM FOR OPTIMAL DESIGN AND DIAGNOSIS OF ELECTRIC POWER DISTRIBUTION GRIDS

DANIEL L. C. E SILVA¹, FELIPE P. B. ZAULI², WELLINGTON F. CANCIAN³, RODNEY R. SALDANHA⁴, DOUGLAS A. G. VIEIRA⁵ AND ADRIANO C. LISBOA⁶

¹ Federal University of Minas Gerais, Brazil  
6627 Avelino Carlos Ave, B. Hte. MG 31270901  
danielsilvamg@ufmg.br

² Federal University of Minas Gerais, Brazil  
6627 Avelino Carlos Ave, B. Hte. MG 31270901  
fzauli@ufmg.br

³ CEMIG Distribuição S.A., Brazil  
1200 Barbacena Ave, B. Hte. MG 30190131  
wfcancian@cemig.com.br

⁴ Federal University of Minas Gerais, Brazil  
6627 Avelino Carlos Ave, B. Hte. MG 31270901  
rodney@cpdee.ufmg.br

⁵ ENACOM - Eng. Assist. por Computador, Brazil  
770 P. J. V. Mendonça St, B. Hte. MG 31310260  
douglas.vieira@enacom.com.br

⁶ ENACOM - Eng. Assist. por Computador, Brazil  
770 P. J. V. Mendonça St, B. Hte. MG 31310260  
adriano.lisboa@enacom.com.br

Key words: Evolutionary Optimization, Statistical Classification, Reliability, Wavelet, Reconfiguration, Power Distribution.

1 INTRODUCTION

Electric utilities usually experience a constant challenge that is intrinsic to their business, characterized by the frequent need for improving their services and, at the same time, reducing the costs for the final consumer [1].

Government and consumers are the main stakeholders in this process. Regulation and frequent evaluations of the quality indexes, along internal objectives of financial and technical efficiency and profitability, significantly influence the company managers strategic decisions.

Within this context, the processes related to power distribution grids maintenance and design represent an important role in order to contribute with the companies towards their goals.

The distribution grids management is an activity that involves the observation and analysis of many inherent aspects, such as reliability, technical and non-technical power losses, maintenance costs, grid topology, among others. Such aspects could be viewed and analyzed
through physical, geographical and, above all, commercial data commonly available in a company’s database.

This paper describes the combined use of mathematical and computational methods for processing data from a simulated power company database and provides a prototype of an auxiliary managing tool, in which it is possible to get a view of the power grid showing its weaknesses and suggesting reconfigurations, in order to minimize such issues.

Figure 1 shows a simplified flowchart which summarizes the prototype operation.

The following text has the following structure: initially, an exposition about the adopted methodology for data extraction, involving $k$-Nearest-Neighbor (kNN) Estimation [2]. The second part describes the techniques used to estimate technical and non-technical losses. The third section demonstrates an approach for estimating service reliability through customer interruption costs. The last methodology section demonstrates the use of Evolutionary Optimization in order to stipulate a re-design for the analyzed network. Finally, the results of a case study are presented followed by the conclusions.

2 METHODOLOGY FOR DATA COLLECTING AND DIAGNOSIS FOR SERVICE RESTORATION COSTS

This study starts with data collection from a simulated power company database: network components geodetic data (feeders, transformers, lampposts) and records of consumer units. The data compilation consisted in performing data mining on technical and commercial databases, integrating the stored information.

Among the main indexes for determining costs, this study has used: service restoration expenses, costs for financial compensation to consumers due to service interruptions, costs of renovation, broken posts, network disruption and other corrective and/or emergency network maintenance services. In Brazil, some of the cited compensations are regulated by [3] and [4].

Using this initial information, the estimation technique $k$-Nearest-Neighbor (kNN) [2] was applied to associate the data of consumer units to the nearest branch from the feeder. As demonstrated in [5], the kNN rule classifies an element $x$ to a category $\theta_n$ if $x_n$ is the nearest neighbor (has the least distance) to $x$ and is also an element of $\theta_n$. 
After grouping the cost data set totalized by consumer unit and bringing together the units by feeder branch (in this case, using kNN with $k=1$), it is possible to develop a graphical representation of each section maintenance cost, allowing the identification of areas that require greater attention of the responsible company.

3 METHODOLOGY FOR TECHNICAL AND NON-TECHNICAL LOSS DIAGNOSIS

3.1 Technical losses

As stated in [1], electrical losses are a natural phenomenon caused by impedance (resistance to the flow of electricity) of electrical devices, such as conductors, transformers, regulators, among others. Although their complete elimination is impossible, they can be minimized through engineering methods.

In this study, the methodology described in [5] is employed to calculate the technical losses through data from consumers history and location, transformers and feeders. The following scheme, adapted from [5], gives an overview about the process:

As described in [5], data from public illumination consists on information about the lamps location, type and power. Transformers information include: location, standard loss specifications and estimated relationship between lifespan and technical losses. The last stage performs the Newton-Raphson method [7] to calculate power flow and estimate customers technical losses.
3.2 Non-technical losses

In a power distribution grid, non-technical losses are mainly caused by electricity theft, in its most common forms, such as fraud, stealing and billing irregularities [8]. In this context, the identification of possible points of theft is a desirable information during the diagnosis process.

Following the methodology used in the previous section, as also described in [5], it is possible to estimate eventual consumer unities with suspicious consumption profile through an image-based analysis using historical data and Wavelet Transform Analysis (WTA).

The WTA here employed is specifically the variation form called Fast Wavelet Transform, in which an image can be decomposed into its horizontal, vertical and diagonal components, allowing the analysis on its specific attributes. Figure 4, adapted from [9], demonstrates and exemplifies its effect.

![Figure 4: Two-dimensional Fast Wavelet Transform analysis and decomposition example.](image)

In this case, periodic images showing the consumer unities’ consumption profile over a map are generated and then processed through WTA. The result is a new image in which is possible to identify points of decadent consumption, which is an evidence for electricity theft.

With the suspicious targets identified and filtered, it is possible to perform a more accurate analysis and inspections to confirm and take the proper measures against the theft situation.
4 METHODOLOGY FOR RELIABILITY DIAGNOSIS

To provide a graphical view showing the reliability of a power distribution grid, the methodology proposed in [10] is used to stipulate reliability through costs originated by violations on the DIC indicator (from Portuguese Duração de Interrupção Individual por Unidade Consumidora), which is provided by Brazilian electric power regulatory agency (ANEEL)[11] and expresses the duration of individual power interruptions for a consumption unity or connection point.

The applied methodology [10] uses the equation (1) to calculate reliability:

\[ Rel_i = \left( 1 - \frac{DIC_i}{DIC_{max}} \right) \times 100 \]  

Where:
- \( Rel_i \): reliability, as percentage, of \( i^{th} \) installation.
- \( DIC_i \): the DIC indicator, in hours, of \( i^{th} \) installation, during the reference month.
- \( DIC_{max} \): the maximum DIC indicator, in hours, stipulated by ANEEL for the region in which the power grid is installed.

As explained in [10], the reliability value, between 0% and 100%, represents the grid’s “slackness degree”. The 0% reliability indicates the necessity of immediate action and a 100% value represents a fully controlled grid.

5 METHODOLOGY FOR GRID RECONFIGURATION

In order to provide another desirable feature into the proposed software prototype, among the explained diagnosis functionalities, this section describes the reconfiguration tool, which completes the proposed set of features.

The basic purpose of such functionality is: after having a diagnostic view of important attributes from the analyzed distribution system, a user would desire to simulate modifications on the grid topology, in order to improve some of its properties.

The following subsections describe an initial version of this feature, which consists in an interactive tool, which allows the user to suggest new branches for the grid and performs simulations through evolutionary optimization to find a new reconfiguration in which the suggested branches can be included, replacing some of the existing ones, or not.

5.1 User’s suggestions input and preparation of the simulation database

As system’s user suggests new branches for the grid, indicating their initial and final points, the next step consists in the estimation of the grid’s new features, needed to the following simulation.

This estimation process starts with the identification of the lampposts that will carry the possible new branches. This is done through a currently rough process in which the closest lampposts are associated to the new branches through kNN rule.

The next step is to estimate the impedance matrix for the new branches. This is made simulating, also through kNN rule, which consumer units will be connected to the each new
branches and calculating each branch’s resistance and reactance through these costumers consumption data.

The last preparative stage is the cycle detection, in which a greedy cycle detection algorithm based in [12] is employed to detect existing cycles on the graph that represents the complete grid (considering existing and suggested branches). This process results in a constraint matrix, which determines how many and which branches can be deactivated, in order to maintain the power grid graph radially connected during the optimization process.

5.3 Simulation for optimal reconfiguration through Evolutionary Optimization

As stated in [13], topology changes in a power distribution grid may be used to promote adjustments and improvements in the system. Among these, it’s highlighted the purpose of reconfiguration for planning and optimal design finalities. Such modifications are possible through the concept of switches, placed at strategic points throughout the system, which connect or disconnect its buses. Figure 5 shows a simplified network scheme, representing a radial distribution system.

Equation 2, based in [13], exposes a condensed mathematical formulation for the grid reconfiguration as an objective function. For this research, it covers only the objective of minimizing technical power losses.
$$\min P_k = \sum_{i=1}^{NR} R_{k,i} \times I_{k,i}^2$$

Where:
- $k$: configuration’s id
- $P_k$: total active losses of $k^{th}$ configuration
- $NR$: number of branches of $k^{th}$ configuration
- $R_{k,i}$: resistance of $i^{th}$ branch of $k^{th}$ configuration
- $I_{k,i}$: current of $i^{th}$ branch of $k^{th}$ configuration

The main objective is subject to the following constraints [13]:
- Radiality
- Tension levels constraints
- Voltage balance constraints between feeders
- Lower and upper bounds of integer and real variables

Each configuration generated by the optimization algorithm is a set of integers, which identifies the buses that could be deactivated for technical power loss reduction. The number of candidate buses for deactivation corresponds to the number of cycles detected in the grid. Thus, it represents a complex combinatorial problem. As affirmed in [14], this kind of problem is called “NP-complete”, as it demands non-deterministic polynomial time to be solved.

There is an expressive number of papers in the literature studying the reconfiguration of distribution networks problem and using optimization techniques. In [15], the use of a heuristic method based on DAOP (discrete ascent optimal programming) is proposed for selection of the best and simplest switching options. In [14], a genetic algorithm aiming at rapid reconfigurations to be employed in service restoration situations was used. Service restoration is also focused in [16], which demonstrates a taboo search method for identifying the necessary switching, while [17] used an ACS type of algorithm (ant colony system) and found promising results analyzing reconfigurations in a planning situation in tests with networks of up to 201 nodes. A multi-objective heuristic was developed in [18]. Finalizing the list of the main consulted references, papers [13], [19] and [20] also tested the use of genetic algorithms.

In the scientific literature on the network reconfiguration problem, the study of meta-heuristic optimization techniques is predominant in the search for optimal solutions. Similarly, the prototype described in the present study employs a meta-heuristic technique based on swarm intelligence known as PSO (Particle Swarm Optimization).

As explained in [21], the PSO algorithm simulates the potential solutions as individual particles in a group moving within the feasible space in a search for the best results. The particles movement is calculated by its initial position changed by its “velocity”, a vector obtained by combining information from the motion of each individual solution with its surroundings information and/or the entire group. A pseudo-code of this procedure is shown below.
a. Generate a random initial group.

b. Repeat
   
   For each particle $i$, do:
   
   Find $p_i$, as the best position of $i$ until now.
   
   Find, in the neighborhood of $i$, $p_g$ as the best-located particle until now.
   
   Calculate the velocity of $i$, considering $p_i$ and $p_g$.
   
   Calculate the new position of $i$.

   End

Until a stop condition is reached.

End.

6 CASE STUDY AND RESULTS

In order to give an example and test the methodology proposed by this paper, a simulated database, representing a miniature of a primary distribution line from a company’s substation feeder was used. It comprises the following data:

- Costumers: 905 consumer units, with geographical data, class.
- Consumption: temporal database with monthly information.
- Transformers: power, initial operation date, geographical data.
- Public Illumination: installation type, power, average loss by model, geographical data.
- Maintenance: costs history, comprising: service restoration, service interruptions, renovation, broken posts, other corrective or emergency services.
- Feeder: geographical data, associated lampposts, data from each bus current state (active power, reactive power), data from branches current state (resistance, reactance).
- Lampposts: geographical data.
- For illustration, background satellite image from Google Maps and map from OpenStreetMap were also used.

Figures 6 to 8 illustrate the prototype’s diagnosis stage. The reconfiguration functionality is illustrated by figures 9 to 11.

![Figure 6: maintenance costs diagnosis. Bluish colors: lower cost. Reddish colors: higher cost.](image)
Figure 7: technical (left) and non-technical losses (right) diagnosis.

Figure 7 shows technical and non-technical losses diagnosis. In the first one, bluish colors indicate lower losses and reddish colors higher losses. On the second, darker dots represent consumption units (targets) with higher electricity theft probability.

Figure 8: reliability diagnosis. Bluish colors: higher reliability. Reddish colors: lower reliability.

Figure 9: Feeder scheme before reconfiguration. Existing branches in black. Suggested branches in red.
7 CONCLUSIONS

The integrated use of kNN, WTA and the reliability estimate method made it possible to process commercial, financial and geographic data to obtain a consistent view of critical points in a distribution network: maintenance costs; technical and non-technical losses; and costs of interruptions.

The implementation of an evolutionary optimization technique to minimize losses was also efficient in fulfilling its objective, as seen in the presented case study. Within the objective of performing the method’s demonstration and its integration with the diagnostic phase of the presented prototype, this optimization approach proved to be a promising candidate for use in
more robust problems, which could involve cost reducing and reliability improving objectives, related with the shown diagnosis.

For continuation of this work, the mathematical model will be expanded in order to add objectives from the presented diagnosis (maintenance costs, non-technical losses and reliability). Improvements in intrinsic computational aspects, such as graph processing, are also in the research interest, as well as the implementation of reconfiguration suggestions generated by the software, through computational intelligence, and the study of a more efficient power flow method, described in [22]. Additionally, for further testing with higher quality, an upgrade in the database will provide a larger amount of information and new physical data about the environment where the distribution grid is installed, as street network and obstacles, among others.

Finally, given the current scenario of study, the built prototype achieved promising results, desirable to a Decision Support System application, representing a rich opportunity to develop a tool for maintenance costs reduction, improvement of service quality (reliability and losses), strategy and grid management in electric utilities.

REFERENCES

[9] Lee, T-H.H. Wavelet Analysis for Image Processing, Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan, ROC.


DIFFERENTIAL EVOLUTION APPLIED TO MOORING OPTIMIZATION OF OFFSHORE FLOATING SYSTEMS BASED ON AN INTEGRATED DESIGN METHODOLOGY

BRUNO F. MONTEIRO¹,², MAURO H. A. DE LIMA JUNIOR¹, BRUNO M. JACOVAZZO¹,², CARL H. ALBRECHT¹,², BEATRIZ S. L. P. DE LIMA² AND BRENNO P. JACOB¹

¹ LAMCSO/COPPE/UFRJ Rio de Janeiro, RJ, Brazil.
LAMCSO—Laboratory of Computational Methods and Offshore Systems, Departamento de Engenharia Civil, Centro de Tecnologia, Bloco B sala B-101, Cidade Universitária, Ilha do Fundão, Caixa Postal 68.506, 21941-914 Rio de Janeiro, RJ, Brasil.
{brunomonteiro, mauro, carl, breno, bruno}@lamcso.coppe.ufrj.br / http://www.lamcso.coppe.ufrj.br

² POLI/UFRJ Rio de Janeiro, RJ, Brazil.
Escola Politécnica - Universidade Federal do Rio de Janeiro
{bruno.monteiro, carl, beatriz, brunoj covazzo}@poli.ufrj.br / http://www.poli.ufrj.br/

Key words: Mooring System, Differential Evolution, SAFOP.

Abstract. The offshore oil production industry is currently expanding its activities in even deeper waters, using moored floating platforms submitted to operational and extreme environmental conditions. Therefore, the design of mooring systems to keep the position of such platforms is of vital importance to assure safety and economical feasibility for offshore activities. Moreover, recently an integrated methodology for the design of mooring systems and risers of floating production platforms has been proposed [1], indicating that the design of the mooring system should consider the riser integrity, by generating a diagram with the safe operational zone (SAFOP) for the risers (comprising an envelope of horizontal excursions of the platform where no design criteria for the risers is violated), to be superimposed to a diagram with the offsets of the floating platform. Therefore, a safe operation is assured when this offset diagram (generated performing global analysis under extreme environmental condition of wave, current and wind for each direction) are within the SAFOP diagram. In this context, this work presents an optimization procedure to determine the mooring configuration leading to the minimum offsets under environmental loads, taking radius, azimuth, pretension and material of the mooring lines as design variables. In previous works, the authors have presented such an optimization tool based on evolutionary algorithms for the synthesis and optimization of mooring system; however, the case studies were relatively simple. Now, this paper presents an implementation of a different algorithm - the Differential Evolution method (DE), which is applied to optimize the mooring system of a real-case scenario for an offshore floating platform. Considering that such optimization procedure requires high computational costs, due to the need of nonlinear static and dynamic analyses with Finite
Element models for each candidate solution, it is very suitable to consider the possibility to apply an algorithm which allows smaller populations, therefore significantly reducing computational time. Differential Evolution, a population-based evolutionary algorithm (EA) has these advantages and has been implemented for this process. This algorithm is a practical approach to global numerical optimization where the number of control parameters in DE is very few; three operators are used including mutation, crossover and selection. Results of the application for real-case systems are presented, which indicate that the method is effective.

1 INTRODUCTION

Offshore oil exploitation activities in deep waters are performed by moored floating platforms in which two of the most important components are the riser system, responsible for the fluid transfer, and the mooring system, responsible to keep the platform horizontal motions into the safe operational zone.

Traditionally, the design of mooring lines and risers has been performed separately, by different teams, with few or no consideration to the other component, and little interaction between the design teams. More recently, an integrated methodology for the design of mooring systems and risers of floating production platforms [1] [2] proposes the full integration of the processes and engineering teams associated to the mooring lines and risers, resulting in efficiency and cost gains.

A proposal for an integrated methodology is to consider the safe operation zone for the risers or SAFOP [2] [3]. This SAFOP is a polar diagram defining the envelope of horizontal excursions allowed to the floating unit where no design criterion for the risers is violated. This diagram is obtained by incrementally displacing the top of risers along several directions, and performing global analyses on various positions under environmental loadings.

Thus, mooring system should keep the floating unit within SAFOP limits. This can be verified by computing the offsets diagram, which is an envelope of the maximum platform horizontal offset under several environmental loadings in multiple directions.

More details about generation of both SAFOP and offsets diagrams can be found in [1].

Therefore, the assessment of the mooring system can be made by plotting both SAFOP and offsets diagram in the same graph. Figure 1 illustrates one comparison of SAFOP and offset diagrams, considering eight directions, with excursions in meters.

In this context, the objective of this work is to present an optimization procedure to find a mooring system configuration that maintains the platform within the safe operation zone for the risers. For this purpose, Differential Evolution method is employed, taking radius, angle and material of lines as variables in the optimization process, and considering the maximum line tension allowed by the design standards.
2 DIFFERENTIAL EVOLUTION

Differential Evolution (DE) is a metaheuristic optimization approach that was first proposed by Price and Storn in 1995 for global optimization [1]. DE has been successfully applied to solving benchmark functions [5], complex problems such as system design [6], solution of the linear systems [7] and robot manipulator design [8]. DE has only three control parameters (mutation control parameter $F$, crossover control parameter $CR$ and population size $NP$), but its optimization performance is sensitive to the choice of these control parameters [9].

The main idea of differential evolution is to generate new individuals from an initial sampling of $Np$ individuals performed randomly with a uniform distribution function within the search space $D$, where at each iteration; each individual $X^{g}$ of the $Np$ has three different individuals randomly chosen to application the operators.

**Mutation** process is done with the difference between two of these individuals and multiplied by a constant and added to a third individual, called base vector.

$$ V^{g}_{i} = X^{g}_{i} + F \cdot (X^{g}_{\alpha} - X^{g}_{\beta}) $$

where, $g$ is the current generation, $i$ is the individual index, $V^{g}$ is the new vector or mutant vector, $X^{g}_{i,\alpha} \neq X^{g}_{i,\beta} \neq X^{g}_{i,\gamma} \in \{1,\ldots,Np\}$ are mutually distinct random individuals and $F$ is the scaling factor typically between 0 and 1 that controls the differential mutation process. For each individual in the population a corresponding mutant vector is generated.

In the **crossover** operation a new vector $W^{g}_{i}$, called trial vector, is created through the discrete recombination with a random value $CR \in [0,1]$, according to the $i$-th population vector and its corresponding mutant vector. So, crossover enhances the potential diversity of a population.

![Image of mooring system assessment](image-url)
Bruno F. Monteiro, Mauro H.A. de Lima Jr., Bruno M. Jacovazzo, Carl H. Albrecht, Beatriz S.L.P. de Lima and Breno P. Jacob

\[
W^g_{ij} = \begin{cases} 
V^g_{i,j} & \text{if } \text{rand}(0,1) \leq CR \text{ or } j = j_{\text{rand}}, \\
X^g_{i,j} & \text{otherwise} 
\end{cases} \quad i = 1, \ldots, NP \quad \text{and} \quad j = 1, \ldots, D
\]

Index \( j_{\text{rand}} \) is a randomly chosen integer within the range \([1, NP]\).

Selection operation is done according to the fitness value of the population vector and its corresponding trial vector to decide which vector will survive to the next generation. For a minimization problem, following selection rule is used:

\[
X^{g+1} = \begin{cases} 
W^g & \text{if } f\left(W^g\right) < f\left(X^g\right), \\
X^g & \text{otherwise} 
\end{cases}
\]

3 MODELING THE OPTIMIZATION PROBLEM

The definition of the configuration of a mooring system involves a very large number of variables. To generate an efficient optimization procedure is necessary to choose a subset of those variables which have relevance to the behavior of the platform. Following are described these selected variables for the optimization process. All other mooring system variables that are not used in the optimization process such as hull characteristics, risers, environmental conditions for analysis, etc. is provided by the base model used in the process.

3.1 Free Variables

Mooring Radius

Mooring radius consists in the horizontal projection between a line connection point on the platform and its anchor point, as illustrated in Figure 2.

![Figure 2: Mooring radius](image)

Angle with the x-axis

This parameter is the line angle with x-axis. Figure 3 shows the line 1 angle.
Work medium tension
Due to the asymmetry of the lines, the system can be unbalanced. Thus, work medium tension is used as the basis for calculating the pre tension of the mooring lines in order to achieve the necessary balance to the project.

Line Material
Line material can be characterized by several parameters: outside diameter, tensile modulus, weight and breaking limit. In this work, all these quantities are related to the line outside diameter by a regression, i.e., their values may be computed from an internal database and approximated as follows.

(a) Outside diameter vs dry weight
(b) Outside diameter vs submerged weight
(c) Outside diameter vs axial stiffness
(d) Outside diameter vs minimum breaking load

Figure 4: Line material parameters
3.2 Objective Function

In order to rank each candidate solution, a fitness function is used, comprised by the objective function and a set of constraints that will be described later.

The objective function, defined by Eq. (4), expresses the offsets diagram approaching the SAFOP diagram, considering a safety factor:

\[
f = \frac{\sum_{i=1}^{NEL} \left( \frac{SF \times SAFOP(i) - Offset(i)}{SF \times SAFOP(i)} \right)}{NEL}
\]  

(4)

Where,

- \( i \) is the direction of active environmental loading.
- \( NEL \) is the total number of environmental loadings, at least 8.
- \( SF \) is the safety factor.
- \( SAFOP(i) \) is the SAFOP limit, in meters, in the direction \( i \).
- \( Offset(i) \) is the greatest vessel offset, in meters, in the direction \( i \).

3.3 Constraints

Four constraints are considered. The first one is related to vessel offset, which can’t be greater than SAFOP limit. Eq. (5) represents this constraint.

\[
v_{offset} = \begin{cases} 
\sum_{i=1}^{NEL} |Offset(i) - SAFOP(i)|, & \text{if } Offset(i) > SF \times SAFOP(i) \\
0, & \text{otherwise}
\end{cases}
\]

(5)

Variables are the same as mentioned in the objective function.

Taking into account the mooring lines tension, API RP-2SK [10] recommends that the tension does not exceed 50% of the rupture limit (MBL) when applied to static analysis. Moreover, the lower limit is restricted to 5% of the MBL to static analysis and intact condition [11]. For these purpose, two constraints functions are applied.

\[
v_{T_{min}} = \begin{cases} 
100 \times \left( 0.05 - \frac{T_{min}}{MBL} \right), & \text{if } \frac{T_{min}}{MBL} \leq 0.05 \\
0, & \text{otherwise}
\end{cases}
\]

(6)

\[
v_{T_{max}} = \begin{cases} 
10 \times \left( 0.5 - \frac{T_{max}}{MBL} \right), & \text{if } \frac{T_{max}}{MBL} \geq 0.5 \\
0, & \text{otherwise}
\end{cases}
\]

(7)

Where,

- \( T_{min} \) is the smallest tension among all lines and all analyzed directions.
- \( T_{max} \) is largest tension among all lines and all analyzed directions.
- \( MBL \) is the line minimum breaking load.
Additionally, the polyester rope shouldn’t touch seabed. This constraint is represented by Eq. (8).

\[ \sum_{i=1}^{N_{\text{Lines}}} L_{\text{Seabed}}(i) \]

Where,

- \( i \) is the mooring line.
- \( N_{\text{Lines}} \) is the total number of mooring lines.
- \( L_{\text{Seabed}} \) is the polyester length, in meters, supported on seabed of the line \( i \).

3.4 Fitness function

Eq. (9) describes the final form of the fitness function.

\[ \text{fitness} = \frac{1}{f + v_{\text{offset}} + v_{\text{min}} + v_{\text{max}} + v_{\text{polyester}}} \]  

Since we seek a solution in which no constraint is violated, this optimization problem becomes a problem of maximizing the fitness function.

4 CASE STUDY

4.1 Model Description

Model studied consists in a real model comprised by a floating platform, 16 mooring lines, 4 per corner and 24 risers. The analyzed scenario is on a water depth of 1800 m.

Mooring lines are composed by three segments: chain, polyester and chain, with physical properties presented in Table 1. Angle between lines is 5° in the same corner, pre tension is set to 2000 kN for each line and mooring radius is 4050 m for each line too.

Risers are flexible type and asymmetrically arranged, and have the same physical characteristics.

Figure 5 shows the numerical model representation, where green represents the mooring lines and blue the risers.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Studless Chain</th>
<th>Polyester</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outside Diameter (m)</td>
<td>0.114</td>
<td>0.21</td>
</tr>
<tr>
<td>Poisson Coefficient</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>Axial Stiffness (kN)</td>
<td>919077.1</td>
<td>187129.0</td>
</tr>
<tr>
<td>MBL - Minimum Breaking Load (kN)</td>
<td>12420</td>
<td>12263</td>
</tr>
<tr>
<td>Dry Weight (kN/m)</td>
<td>2.549</td>
<td>0.2845</td>
</tr>
<tr>
<td>Hydrodinamic Diameter (m)</td>
<td>0.2</td>
<td>0.21</td>
</tr>
</tbody>
</table>

4.2 Analysis Parameters

It is well known that, on actual riser/mooring design activities, full non-linear, time-domain dynamic Finite Element may be required to accurately evaluate the behavior of the floating system. However, along the optimization process with the DE algorithm, a very large number of analyses should be performed for all candidate configurations. If such analyses
were performed using full FE dynamic analyses, very high computational costs would be required.

The main objective of this work is to assess the behavior of the DE algorithm associated to the optimization of mooring systems, and to qualitatively illustrate the mooring configurations obtained. Therefore, in the case study presented here the evaluations are performed using static analyses only, which are much faster to compute and provide results that are at least representative of the actual dynamic solution. Of course, on an actual use of the procedure for design activities, the optimization procedure can employ a dynamic solver and rely on its parallel implementation on clusters of computers.

![Model](image1)

**Figure 5:** Model – 3D and top view

### 4.3 Environmental Loads

Considering the use of static analysis as mentioned above, the environmental loads consider current only, acting on coupled models including the risers, mooring lines and the platform hull. Triangular profile is applied in eight directions (N, NE, E, SE, S, SW, W, NW), with intensity at sea surface according to Table 2.

<table>
<thead>
<tr>
<th>Direction</th>
<th>Velocity at sea surface (m/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>1.30</td>
</tr>
<tr>
<td>NE</td>
<td>1.17</td>
</tr>
<tr>
<td>E</td>
<td>0.91</td>
</tr>
<tr>
<td>SE</td>
<td>1.26</td>
</tr>
<tr>
<td>S</td>
<td>1.39</td>
</tr>
<tr>
<td>SW</td>
<td>1.26</td>
</tr>
<tr>
<td>W</td>
<td>1.18</td>
</tr>
<tr>
<td>NW</td>
<td>1.17</td>
</tr>
</tbody>
</table>

### 4.4 SAFOP

SAFOP limits for each direction can be seen in the Table 3.
Table 3: SAFOP limits

<table>
<thead>
<tr>
<th>Direction</th>
<th>Limit (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>253.13</td>
</tr>
<tr>
<td>NE</td>
<td>210.94</td>
</tr>
<tr>
<td>E</td>
<td>140.63</td>
</tr>
<tr>
<td>SE</td>
<td>154.69</td>
</tr>
<tr>
<td>S</td>
<td>210.94</td>
</tr>
<tr>
<td>SW</td>
<td>281.25</td>
</tr>
<tr>
<td>W</td>
<td>182.81</td>
</tr>
<tr>
<td>NW</td>
<td>225.00</td>
</tr>
</tbody>
</table>

4.5 Design Variables

According to section 3.1, Table 4 shows design variables for the optimization process of this case study. It should be noted that the variables angle and mooring radius have a variation surrounding the original model.

Table 4: Physical properties of Mooring Lines (Original Model)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Upper bound</th>
<th>Lower bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-4</td>
<td>Angle per corner</td>
<td>-3º</td>
<td>3º</td>
</tr>
<tr>
<td>5-8</td>
<td>Radius per corner</td>
<td>-500 m</td>
<td>500 m</td>
</tr>
<tr>
<td>9</td>
<td>Medium tension</td>
<td>2000 kN</td>
<td>5000 kN</td>
</tr>
<tr>
<td>10</td>
<td>Material (diameter)</td>
<td>0.122 m</td>
<td>0.262 m</td>
</tr>
</tbody>
</table>

4.6 Algorithm Parameters

Ten independents executions were performed. Population size was set to $NP = 6$ individuals for each generation. DE coefficients of mutation and crossover were set to $F = 0.8$ and $CR = 0.2$, respectively.

Stopping criterion used was 120 fitness evaluations.

5 RESULTS

5.1 Algorithm Evolution

This section presents the application results of the optimization procedure. Figure 6 shows the global best evolution for the best execution, the global best for the worst execution, average and standard deviation for the global best among 10 executions for each fitness evaluation.

As we can see, the global best for the best execution evolves quickly. Note also that there is a stagnation of fitness value; it shows the number of fitness evaluations was satisfactory.
Figure 6: Global best evolution during the fitness evaluations

5.2 Optimized Model

Figure 7 shows a 3D and top view of the optimized model allowing a visual comparison between this one and original model represented in Figure 5.

![Optimized Model – 3D and top view]

Optimized material had the outside diameter equal to 0.216 m. Other parameters are calculated from this, for example, MBL is equal to 17670.03 kN.

From Figure 7 and Table 5, it is noted that the mooring radius of the optimized model is smaller than original one. This means a more compact field arrangement, which enables deployment in nearby areas, other units of production / drilling / intervention that also use mooring systems, offloading pipelines and other subsea components.

Table 6 shows the optimized offset range d from 38.5% to 79.0% of the external envelope of SAFOP, which also meets the established safety limit of 20%, i.e., 80% of SAFOP limit. On the other hand, the offsets of original model violate even SAFOP, which does not guarantee the risers integrity.

Figure 8 shows the superposition of SAFOP and offsets diagrams of both original and optimized models enabling a visual assessment of mooring system efficiency.

Table 6 also shows lines tensions are within the limits set by the rules.
Table 5: Comparative of angles and mooring radii between original and optimized models

<table>
<thead>
<tr>
<th>Line</th>
<th>Angle (%)</th>
<th>Radius (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original</td>
<td>Optimized</td>
</tr>
<tr>
<td>1</td>
<td>42.0</td>
<td>41.4</td>
</tr>
<tr>
<td>2</td>
<td>47.0</td>
<td>46.4</td>
</tr>
<tr>
<td>3</td>
<td>52.0</td>
<td>51.4</td>
</tr>
<tr>
<td>4</td>
<td>57.0</td>
<td>56.4</td>
</tr>
<tr>
<td>5</td>
<td>127.5</td>
<td>125.6</td>
</tr>
<tr>
<td>6</td>
<td>132.5</td>
<td>130.6</td>
</tr>
<tr>
<td>7</td>
<td>137.5</td>
<td>135.6</td>
</tr>
<tr>
<td>8</td>
<td>142.5</td>
<td>140.6</td>
</tr>
<tr>
<td>9</td>
<td>226.0</td>
<td>228.5</td>
</tr>
<tr>
<td>10</td>
<td>231.0</td>
<td>233.5</td>
</tr>
<tr>
<td>11</td>
<td>236.0</td>
<td>238.5</td>
</tr>
<tr>
<td>12</td>
<td>241.0</td>
<td>243.5</td>
</tr>
<tr>
<td>13</td>
<td>300.0</td>
<td>300.2</td>
</tr>
<tr>
<td>14</td>
<td>305.0</td>
<td>305.2</td>
</tr>
<tr>
<td>15</td>
<td>310.0</td>
<td>310.2</td>
</tr>
<tr>
<td>16</td>
<td>315.0</td>
<td>315.2</td>
</tr>
</tbody>
</table>

Table 6: Comparative of offsets and lines tension between original and optimized models

<table>
<thead>
<tr>
<th>Direction</th>
<th>Offset (m)</th>
<th>Tension</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original</td>
<td>Optimized</td>
</tr>
<tr>
<td></td>
<td>(m)</td>
<td>(kN)</td>
</tr>
<tr>
<td>N</td>
<td>55.8</td>
<td>97.5</td>
</tr>
<tr>
<td>NE</td>
<td>119.0</td>
<td>136.4</td>
</tr>
<tr>
<td>E</td>
<td>151.3</td>
<td>111.1</td>
</tr>
<tr>
<td>SE</td>
<td>182.7</td>
<td>109.4</td>
</tr>
<tr>
<td>S</td>
<td>167.7</td>
<td>135.5</td>
</tr>
<tr>
<td>SW</td>
<td>155.2</td>
<td>163.2</td>
</tr>
<tr>
<td>W</td>
<td>67.4</td>
<td>132.6</td>
</tr>
<tr>
<td>NW</td>
<td>30.2</td>
<td>145.8</td>
</tr>
</tbody>
</table>

Figure 8: Superposition of SAFOP and Offsets diagrams
6 CONCLUSIONS

This work describes the modeling and implementation of a computational tool for the optimization of mooring systems for floating production platforms considering an integrated methodology for the design of mooring systems and risers. Thus, we can define as objective of the optimization process the integrity of risers system, in order to attend SAFOP.

The case study presented the application of the optimization tool to scenarios representative of real case using Differential Evolution algorithm. Results indicated the optimization process provides automatically from a base model, a mooring system configuration in which vessel offsets assures the integrity of the risers maintaining lines tensions within the limits established by the design rules. Optimized model also has a smaller mooring radius than an original model, providing a more compact field arrangement.

As mentioned before, for its actual use on design activities, dynamics analyses should be incorporated in the optimization process. Since this increases significantly the computational cost, current developments are focusing on reducing the number of evaluations and employing parallel implementations, allowing the practical use of this procedure.

REFERENCES


ON THE PERFORMANCE OF THE HARMONY SEARCH ALGORITHM IN THE OPTIMIZATION OF LAMINATED COMPOSITE PLATES

FELIPE S. DE ALMEIDA

Department of Civil Engineering, Federal University of Rio Grande do Sul
Av. Osvaldo Aranha 99 (90135-190), Porto Alegre, RS, Brasil
felipe.almeida@ufrgs.br

Key words: Harmony search algorithm, laminated composite plates, discrete optimization

Abstract. This paper presents an evaluation of the performance of a relatively new heuristic method called Harmony Search (HS) algorithm in the optimization of the stacking sequence of a laminated plate. The objective of the optimization is to maximize the critical buckling load of a symmetric and balanced laminated plate containing 32, 48 or 64 plies with fiber orientations of $0^\circ$, $\pm 45^\circ$ or $90^\circ$. The performance of the algorithm is determined based on the apparent reliability, which is used to evaluate the sensitivity of method with respect to the parameters harmony memory size (HMS), harmony memory consideration rate (HMCR) and pitch adjusting rate (PAR). Tests show that HMS has strong influence on HS performance and the best results are obtained with the higher HMS values. Less sensitivity was observed with respect to HMCR, but the PAR has a considerable effect on the apparent reliability. High levels of R were obtained with PAR values between 0.1 and 0.3. The last study investigate the effects of a linear variation of PAR during the optimization. Results indicate that PAR should be increased as the optimization advance since good reliability levels and less sensitivity to this parameter are obtained with this strategy.

1 INTRODUCTION

Unidirectional fiber reinforced composites are interesting materials because of its high specific strength in the fiber direction. These materials are used in laminated plates where each ply may have a different fiber orientation. The fact that the laminate stacking sequence can be tailored to a specific geometric and load configuration is one of the main advantages of using such materials. This process is accomplished by taking the fiber orientation of each one of the laminate plies as a design variable that can assume only few discrete values such as $0^\circ$, $\pm 45^\circ$ or $90^\circ$. Obtaining the optimum configuration for a laminate becomes a harder task as the number of plies increase, which make the use of optimization methods suitable for the design of laminated composite structures.
This work presents the evaluation of the performance of a relatively new optimization method called Harmony Search Algorithm (HSA) [1] in the optimization of the staking sequence of laminated composite plates. HSA has been successfully applied in the optimization of many other engineering problems, but only two works about the optimization of laminate structures were published [2, 3]. Furthermore, these works lack of a consistent evaluation of the algorithm performance and its sensitivity to parameters setting, which is the main objective of the research presented in the following sections.

2 HARMONY SEARCH ALGORITHM AND ITS VARIANTS

Harmony search (HS) is a metaheuristic algorithm inspired on the process of improvisation in music playing. The design variables of a new solution vector are defined by random, simulating the improvisation in the selection of a musical note, or copied from a pool of the best previous solution vectors, which correspond to the musician experience in the melody. There is a small chance that the selected value is replaced by a neighbouring value, which represents the pitch adjustment in music improvisation.

Solution vectors are represented by a harmony vector $\mathbf{x} = \{x_1, x_2, x_3, \ldots, x_{nv}\}$, where $nv$ is the number of design variables and $x_i \in [L_i, U_i]$ is the value of the design variable $i$, with lower bound and upper bound $L_i$ and $U_i$, respectively. The best solution vector obtained during the optimization are stored in the harmony memory HM and used for the creation of new solution vectors. The number of solution vectors stored in HM is defined by the parameter harmony memory size HMS.

The HS algorithm is usually described in literature [1, 4] by the following steps:

Step 1 – Initialization: Definition of the objective function $f(\mathbf{x})$, the number of design variables $nv$, the lower bound ($L_i$) and upper bound ($U_i$) of each design variable $i$ and the value of the HS parameters.

Step 2 – Initialize the harmony memory: The HM is filled with $HMS$ solution vectors that are randomly generated.

Step 3 – Improvise a new harmony: A new solution vector is created by the process defined in the algorithm 1. The value of each design variable $x_i^{\text{new}}$ is selected form the HM with a chance given by the parameter HMCR (HM consideration) or it is randomly defined with a chance 1-HMCR. The pitch adjustment may be applied with a chance defined by the parameter pitch adjusting rate (PAR) when the value of the variable is selected from HM. In this case a neighbouring value is chosen for the variable in a range defined by the parameter random distance bandwidth ($bw$). The objective function is calculated after the creation of the new solution vector. In this paper, $\text{irand}(c,d)$ represents a random integer in the range $[c,d]$ and $\text{rand}(a,b)$ represents a real random number uniformly distributed in the interval $[a,b]$.

Step 4 – Update the harmony memory: If the new vector ($\mathbf{x}^{\text{new}}$) is better than the worst vector in the HM in terms of the objective value, $\mathbf{x}^{\text{new}}$ is stored replacing the worst vector in HM.

Step 5 – Check the stopping criterion: The vector with the best objective value in the HM is taken as the solution of the optimization process if the stopping criteria is satisfied. Otherwise, the number of iterations ($t$) is incremented and steps 3 and 4 are repeated. In this work, the
Algorithm 1Improvisation of a harmony vector $x^{new}$

1: for $i=1,nv$ do
2: \hspace{1em} if $\text{rand}(0,1) \leq \text{HMCR}$ then \hspace{1em} $\triangleright$ HM consideration
3: \hspace{2.5em} $j \leftarrow \text{irand}(1,\text{HMS})$
4: \hspace{2.5em} $x^{new}_i \leftarrow x^j_i$
5: \hspace{1em} if $\text{rand}(0,1) \leq \text{PAR}$ then \hspace{1em} $\triangleright$ Pitch adjustment
6: \hspace{2.5em} if $\text{rand}(0,1) \leq 0.5$ then
7: \hspace{4em} $x^{new}_i \leftarrow \max(L_i, x^{new}_i - \text{irand}(1,bw))$
8: \hspace{2.5em} else
9: \hspace{3.5em} $x^{new}_i \leftarrow \min(U_i, x^{new}_i + \text{irand}(1,bw))$
10: \hspace{2.5em} end if
11: \hspace{1em} end if
12: else \hspace{1em} $\triangleright$ ($x^{new}_i$ is randomly chosen)
13: \hspace{2.5em} $x^{new}_i \leftarrow \text{irand}(L_i,U_i)$
14: end if
15: end for
16: Calculate $f(x^{new})$

search terminates when the number of objective function evaluations ($E$), corresponding to the number of different solution vectors generated during the optimization process, reaches the maximum value $E_{max}$, which is a parameter defined for each problem.

2.1 Variants of the HS algorithm

A comprehensive overview of the modification proposed in order to improve the performance of the original HS is given in [5]. The first variant of HS, called Improved Harmony Search (IHS), was introduced by [4]. The modification incorporated in IHS is the concept of the continuous variation of the parameter $\text{PAR}$ and $bw$ during the optimization. The evolution of $\text{PAR}$ with respect to the current algorithm iteration ($t$) is given by

$$\text{PAR}(t) = \text{PAR}^I + \frac{\text{PAR}^F - \text{PAR}^I}{t_{max}} \cdot t$$

(1)

where $\text{PAR}^I$ and $\text{PAR}^F$ are the initial and final value of $\text{PAR}$, respectively. In IHS, $\text{PAR}$ is linearly increased during the optimization with $\text{PAR}^I$ and $\text{PAR}^F$ being equal to the minimum and maximum value, respectively. A different approach is adopted in [6] by decreasing $\text{PAR}$ during the optimization, which is accomplished with Eq. 1 using $\text{PAR}^I$ and $\text{PAR}^F$ equal to the maximum and minimum value of $\text{PAR}$, respectively.

This work proposes and tests a small modification to the basic pitch adjustment given in the algorithm 2, which replaces the lines 6 to 10 in the algorithm 1.
Algorithm 2 Modified pitch adjustment

1: if \( x_{i}^{\text{new}} = L_{i} \) then
2: \( x_{i}^{\text{new}} \leftarrow x_{i}^{\text{new}} + \text{irand}(1,bw) \)
3: else if \( x_{i}^{\text{new}} = U_{i} \) then
4: \( x_{i}^{\text{new}} \leftarrow x_{i}^{\text{new}} - \text{irand}(1,bw) \)
5: else
6: if \( \text{rand}(0,1) \leq 0.5 \) then
7: \( x_{i}^{\text{new}} \leftarrow \max(L_{i},x_{i}^{\text{new}} - \text{irand}(1,bw)) \)
8: else
9: \( x_{i}^{\text{new}} \leftarrow \min(U_{i},x_{i}^{\text{new}} + \text{irand}(1,bw)) \)
10: end if
11: end if

3 FORMULATION OF THE OPTIMIZATION PROBLEM

This work studies the optimization of laminate stacking sequence in order to maximize the critical buckling load of the laminated plate shown in Fig. 1. Only symmetric and balanced stacking sequences are considered and the laminate is made of a fiber reinforced material with principal direction given by the axis 1 and 2, respectively. The angle \( \theta_{i} \) defines the orientation of the axis 1 of each ply \( i \). A limit of four contiguous plies with the same fiber orientation is imposed in order to prevent the propagation of cracks in the matrix though the laminate thickness.

The buckling load factor \( \lambda_{b} \) corresponding to a buckling mode with \( m \) and \( n \) half waves in the \( x \) and \( y \) directions, respectively, is given by [7]

\[
\lambda_{b}(m,n) = \frac{\pi^{2}}{2} \left[ \frac{m^{4}D_{11} + 2(D_{12} + 2D_{66})(rnm)^{2} + (rn)^{4}D_{22}}{(am)^{2}N_{x} + (ran)^{2}N_{y}} \right]
\] (2)

where \( D_{ij} \) are the components of the laminate bending stiffness and \( r = a/b \) is plate aspect ratio. The buckling load given by Eq. 2 is approximately exact for laminates formed by plies with angles \( \pm \theta \) in sequence because the terms \( D_{16} \) and \( D_{26} \) are neglected [7]. The critical buckling load

![Figure 1: Laminated plate under biaxial compressive load](image-url)
load \( N_{x,cb} = \lambda_{cb} N_x \) and \( N_{y,cb} = \lambda_{cb} N_y \) is determined by the critical buckling load factor \( \lambda_{cb} \), which is the minimum value of \( \lambda_b \) obtained over all possible combinations of the number of half waves \( m \) and \( n \).

The objective function to be minimized in the optimization is defined as

\[
f(x) = -\lambda_{cb}(\theta)
\]

where \( \theta \) is the vector containing the angles \( \theta_i \) of the laminate stacking sequence, which are obtained by the decodification of the solution vector \( x \) generated in each iteration of the HS algorithm.

This works adopts the integers 1, 2 and 3 for the value of the design variables to represent the plies \( 0_2, \pm 45 \) and \( 90_2 \), respectively. Only half of the staking sequence is defined by the design variables for the symmetric laminate, with the position \( x_1 \) and \( x_{nv} \) associated to the outermost plies and the innermost plies, respectively. A solution vector with \( nv \) design variables represents a laminate with a number of plies \( N_{plies} \) equal to \( 4 \times nv \).

A special strategy is adopted to prevent stacking sequences containing blocks of \( 0_n \) or \( 90_n \) plies that violate the constraint on the maximum number of contiguous plies with the same fiber orientation, which is limited to 4 in this work. The first variable \( x_i, i = 1, 2, \cdots, nv \) that lead to a violation of this constraint is converted into a \( \pm 45 \) ply stack.

The research presented in this paper considers three optimization cases (OC1, OC2 and OC3) related to the maximization of the buckling load of the laminated plate shown in Fig. 1. These cases differ from each other only in terms of the number of plies of the laminate \( (N_{plies}) \), which is reflected in the number of design variables \( (nv) \) that is given in Tab. 1. The plate have simple supported sides of length \( a = 0.5 \text{ m} \) and \( b = 1 \text{ m} \). The in-plane compressive load are \( N_x = 0.333 \text{ N/m} \) and \( N_y = 1 \text{ N/m} \). All plies have the same thickness of 0.127 mm. The mechanical properties with respect to the material principal directions 1–2 are: Young modulus \( E_1 = 127.6 \text{ GPa} \) and \( E_2 = 13 \text{ GPa} \), in-plane shear modulus \( G_{12} = 6.4 \text{ GPa} \) and Poisson’s ratio \( \nu_{12} = 0.3 \).

Table 1 also shows the size of the design space (SDS), corresponding to the total number of possible solution vectors, the number of optimum solution vectors \( (N_{ot}) \) and the corresponding critical buckling load factors \( (\lambda_{cb}^{ot}) \) for each case investigated. The number of variables adopted resulted in a small design space for OC1, a medium design space for OC2 and a large design space for OC3.

<table>
<thead>
<tr>
<th>Case</th>
<th>( nv )</th>
<th>( N_{plies} )</th>
<th>SDS</th>
<th>( N_{ot} )</th>
<th>( \lambda_{cb}^{ot} )</th>
<th>( E_{max} )</th>
<th>( E_{max}/SDS )</th>
</tr>
</thead>
<tbody>
<tr>
<td>OC1</td>
<td>8</td>
<td>32</td>
<td>6561</td>
<td>1</td>
<td>41250</td>
<td>1000</td>
<td>15.24%</td>
</tr>
<tr>
<td>OC2</td>
<td>12</td>
<td>48</td>
<td>531441</td>
<td>7</td>
<td>139348</td>
<td>5000</td>
<td>0.941%</td>
</tr>
<tr>
<td>OC3</td>
<td>16</td>
<td>64</td>
<td>43046721</td>
<td>50</td>
<td>330326</td>
<td>15000</td>
<td>0.035%</td>
</tr>
</tbody>
</table>
4 RESULTS AND DISCUSSIONS

The performance of the HS algorithm is evaluated based on the apparent reliability ($R$) of the method, which is defined as the chance to find a global optimum in a single search. Results are presented in graphs relating $R$ to the number of objective function evaluations in a single run of the HSA ($E$). The apparent reliability is obtained by the expression

$$R(E) = \frac{M_{opt}(E)}{M}$$

where $M_{opt}(E)$ is the number of optimizations that obtained a global optimum with less than $E$ objective function evaluations and $M$ is the number of runs of the algorithm for each test, which is adopted as 200 runs in this work.

The amount of computational work, defined by the parameter $E_{max}$, allocated for the optimizations is given in Tab. 1. The computational workload of OC3 is 15 times the workload of OC1 but the portion of the possible design vectors explored in the former (given by $E_{max}/SDS$) is much smaller.

4.1 Results for the Basic HS

This section presents the evaluation of the performance of the Basic HSA in the problems defined in section 3. The Basic HSA is described in the initial part of section 2, with addition to the modification given in the algorithm 2. All HS parameters have static values in the tests reported below.

Section 4.1.1 is dedicated to the presentation of the results obtained in the OC2. A more refined investigation about the influence of the HS parameters on the algorithm performance is conducted for this case. A study about the impact of the modification on the pitch adjustment process, proposed in the algorithm 2, is also developed. Section 4.1.2 presents the results obtained in the OC1 and OC3. The sensitivity of the HS performance with respect to $nv$ or SDS is discussed based on the results of the three cases.

4.1.1 Tests of optimization case 2

The HSA is applied in the OC2 using HMS of 12, 24 or 48 design vectors, which corresponds to ratios $HMS/nv$ equal to 1, 2 and 4, respectively. The optimization were conducted with HMCR equal to 0.90, 0.95 or 0.99 and PAR equal to 0.1, 0.2, 0.3, 0.4 or 0.5.

The graphs relating $R$ and $E$ are presented in figures 2a, 2b and 2c for the optimizations using HMCR equal to 0.90, 0.95 and 0.99, respectively. These results show that HMCR is the parameter with less influence on the performance of the algorithm.

Generally, the best R levels in the small E range ($E < 1000$) are obtained with lower values of PAR and HMS. As more points of the design space are explored in the search ($1000 < E < 4000$), the optimizations with larger memory (higher HMS) and $0.1 \leq PAR \leq 0.3$ presented better R levels. For $E = E_{max}$, most of the parameter configurations resulted in R greater than 90%. The exception are the optimizations with PAR equal to 0.5 and the cases with low values
of HMS and PAR, for which the reliability do not increase with larger searches. The combination of small HMS and PAR equal to 0.1 resulted in the best reliability ($R \approx 0.4$) if very few function evaluations are allowed in the optimization ($E_{\text{max}} \leq 400$).

Figure 2: Apparent reliability in OC2

An additional study is performed for OC2 in order to evaluate the effect of the modified pitch adjustment (see algorithm 2) on the apparent reliability. The original version of HS, which is
given in algorithm 1, is applied in the OC2 with the HMCR equal to 0.95. The performance for this case is shown in Fig. 3 where the apparent reliability is considerably reduced if compared to the results of the HS using the modified pitch adjustment in Fig. 2b.

Figure 3: Apparent reliability in OC2 with the original pitch adjustment

4.1.2 Tests of optimization cases 1 and 3

The application of the basic HSA to the OC1 and OC3 is presented in this section. The tests are performed considering the HMCR equal to 0.99 and HMS equal to 2 or 4 time the number of variables in each case.

The apparent reliability obtained in the OC1 using PAR equal to 0.1, 0.2, 0.3, 0.4 or 0.5 are presented in Fig. 4a. This optimization case has a very small design space when compared to OC2, but a much larger portion of the design space is explored in each search (see $E_{max}/SDS$ in Tab. 1). The curves of $R \times E$ are very close to each other, indicating that PAR has less influence on the HSA performance in OC1 than in OC2. The best results are obtained with $0.2 \leq PAR \leq 0.4$, which is different form the best range of values observed for OC2. As observed in OC2, optimizations using less memory (low values of HMS) perform better in the early stages of the search while optimizations using more memory show the best results when $E = E_{max}$. HS have to explore at least 10% of the design space ($E \approx 650$) to reach a reliability level of 80% with the best parameter sets, but levels higher than 90% are obtained with many combination of parameters when the stopping criterion is satisfied ($E = E_{max} = 1000$).

Results obtained in OC3 are presented in the remainder of this section. Figure 4b shows the apparent reliability of the HSA using PAR equal to 0.05, 0.1, 0.2 or 0.3. Even with much more function evaluations than in OC2, only a tiny part of the more than 43 million possible solution vectors in OC3 are searched in each optimization process. The parameter PAR have a strong influence on the HSA performance in OC3. Optimizations using PAR equal to 0.1 and 0.2 have better apparent reliability for $E < 2E_{max}/3$, but $R > 90\%$ is guaranteed only at the end of the search in optimizations using PAR equal to 0.3. Larger memories are clearly beneficial for this optimization case with a large number of design variables.
4.2 Results for HS with varying PAR

This section presents the performance of two variations of the HSA, which differ from the basic HSA by considering a linear variation of PAR with respect to the iterations of the algorithm. Section 4.2.1 shows the results for the optimizations with PAR being increased during the optimization, as proposed in the Improved Harmony Search (IHS) [4]. The followin section shows the results for an alternative strategy [6] where PAR is decreased during the optimization.

4.2.1 Tests with increasing value of PAR

Results of the apparent reliability of HSA with PAR increased during the optimization are presented in figures 5a, 5b and 5c, for the OC1, OC2 and OC3, respectively. Each optimization case is studied with HMS coinciding with the values used in section 4.1 for an easy comparison with the basic HSA. The pitch adjustment is applied with a chance PAR given by Eq. 1 with \( PAR^l \) equal to 0.05 or 0.1 and \( PAR^F \) equal to 0.2, 0.3 or 0.4.

All test resulted in curves \( R \times E \) that are very close in the early stages of the optimizations for all PAR values and show a small dispersion as E approaches \( E_{max} \). The low sensitivity with respect to the parameters \( PAR^l \) and \( PAR^F \) is a positive aspect of the algorithm investigated in this section.
The apparent reliability obtained in OC1 is similar to the reliability levels in section 4.1 (see figures 4a and 5a). The main difference obtained with the strategy used in this section is the elimination of the pronounced plateau in the curve $R \times E$ for PAR equal to 0.1 that is observed for the basic HSA. Figure 5b shows that in OC2 a fast increase of R occur in the initial range of E and high reliability levels for all combinations of $PARI$ and $PARF$ are obtained for $E$ equal to $E_{max}$. The best results were obtained with $PARI$ equal to 0.10, for which the reliability level is very close to the best results obtained with the basic HSA (see Fig. 2c).

The apparent reliability obtained in OC3 (see Fig. 5c) is intermediate with respect to the higher and lower reliability levels observed for de basic HSA in Fig. 4b, but almost the same results are obtained in both cases at the end of the optimization process.

### 4.2.2 Tests with decreasing value of PAR

The performance of HSA with PAR decreased during the optimization is presented in figures 6a, 6b and 6c for the OC1, OC2 and OC3, respectively. The value of the parameter PAR is given by Eq. 1 with $PARI$ equal to 0.2, 0.3 or 0.4 and $PARF$ equal to 0.05 or 0.1. The best results in OC1 are obtained with $PARI = 0.4$, while $PARI = 0.2$ is the best value for the OC2 and OC3.
The apparent reliability obtained in OC1 with the version of HSA studied in this section have maximum values that are close to the best results presented in the previous sections. However, Fig. 6a shows a big difference in the curves $R \times E$ at the final part of the optimization, indicating a strong influence of the parameters $PAR^I$ and $PAR^F$.

The highest reliability levels obtained in OC2 are better or similar to the best results obtained for the same case in the previous sections. All tests shown in Fig. 6b obtained $R > 90\%$ with less than 3000 function evaluations and optimizations using HMS equal 48 obtained $R \geq 99\%$ at the end of the search. On the other hand, a pronounced degradation of $R$ is observed in the first stage of the optimization depending on the parameters $PAR^I$ and $PAR^F$.

In spite of presenting excellent results at the end of the search in OC3, the algorithm adopting a progressive reduction of PAR performs much worse than the previous versions of the HSA for intermediate range of $E$ in this case. This is a very important characteristic since the minimal number of function evaluations that are necessary to attain an appropriate reliability level in a specific problem is not known a priori or may not be allowed due to restraints on the computational cost.

Figure 6: Apparent reliability with HMCR=0.99 and decreasing PAR value
5 CONCLUSIONS

Results reported in previous sections showed that HS is able to find the global optimum in laminate stacking sequence optimizations, but the modified pitch adjustment proposed in algorithm 2 is essential to improve the performance of the method in this kind of problem.

Good reliability levels are obtained with the basic HSA using all values of HMCR and the highest HMS tested. However, a strong sensitivity to PAR is observed and this parameter has to be properly set in the range (0,0.3] for each problem in order to achieve the best performance.

The variant of HSA considering an increasing value of PAR during the optimization performed well in the problems studied in this work. The reduced sensitivity to the parameters $\text{PAR}^d$ and $\text{PAR}^f$ is the main advantage of this strategy. This behaviour is associated with good reliability levels in the first stage of the optimization, which is particularly interesting since few function evaluations use the be allowed in the optimization of practical engineering problems.

The algorithm that adopted a reduction of PAR during the optimization obtained excellent results at the end of the search in the OC2 and OC3. However, results indicate that the method may perform worse than the other versions of HS if the number of function evaluations is moderately reduced, which is a drawback for problems with costly objective function calculation.

REFERENCES


PARALLEL GENETIC ALGORITHMS FOR THE DESIGN OF MIMO CONTROL SYSTEMS

PAOLA P. OTEIZA¹, 2, DIEGO A. RODRIGUEZ¹, 2, 3, NÉLIDA B. BRIGNOLE¹, 2

¹Planta Piloto de Ingeniería Química (PLAPIQUI), CCT-UAT, UNS-CONICET,
Camino La Carrindanga Km. 7, Bahía Blanca, Buenos Aires
poteiza@plapiqui.edu.ar

²Laboratorio de Investigación y Desarrollo en Computación Científica (LIDeCC), DCIC-UNS,
Av. Alem 1253, Bahía Blanca, Buenos Aires.
dybrigno@criba.edu.ar

³Facultad de Ciencias Exactas, Universidad Nacional de Salta (UNSa)
Av. Bolivia 5150, Salta, Salta.
drodriguez@plapiqui.edu.ar

Keywords: parallel processing; Genetic Algorithms, MIMO, control, metaheuristics,

Abstract. In this work a metaheuristic technique for unconventional MIMO controller design was explored and its performance was assessed for a real-world case. We aimed at finding the optimal parameters in order to achieve a satisfactory control for the volume and output of natural gasoline in a storage tank. By means of the implementation of parallel algorithms, both the computational time and programming effort could considerably be reduced and the solutions were improved. The preliminary GA results were obtained with Gaussian mutations combined with various crossovers. The experimental evaluation yielded satisfactory fitness values. By means of parallel programming, relatively little computational effort was required, especially with a scattered crossover.

1. INTRODUCTION

Before the decade of the 1940s, most of the industrial processes were controlled either manually or with very primitive controllers. Concomitantly, the extensive industrialization phenomenon required higher production volumes. In consequence, the amount of variables to be controlled in a plant increased significantly. Automatic control became an attractive solution of plant-operability problems. Therefore, it grew considerably since 1950.

Most of the systems are multivariable and exhibit interaction among their variables. The design of a control system should take this phenomenon into account so as to guarantee reliable regulation. Inspired in the field of numerical analysis, Jyh-Cheng Jeng [1] presented a novel closed-loop identification method for multivariable interactive systems from step responses, where he used both B-spline series expansions and a linear least-squares technique.

At the end of the 20th century, Evolutionary Algorithms (EAs) grew spectacularly. Their practical application extended widely to many knowledge areas, where the diverse problems had a common feature: they were difficult or impossible to solve with traditional methods. In the fields of control and multiobjective optimization, EAs were helpful to find a relatively good solution. In particular, at the onset of the 21st century, Viswanathan et al. [2] successfully addressed the closed-loop identification problem of two-input two-output (TITO) processes and they employed a Genetic Algorithm (GA) to locate reliably the global minimum of the least-squares problem. Later, Wei-Der Chang [3] proposed a modified
crossover formula in GAs, and used this method to determine PID controller gains for multivariable processes. Besides, Meng and Song [4] presented fast GAs, exhibiting improvements about population, selection, crossover and mutation in comparison with simple GAs. In turn, as to self-tuning systems, Won Seok Oh et al. [5] proposed a genetic-based self-tuning speed controller for the high-performance drives of induction motors. In contrast, Liu and Hsu [6] designed a self-tuning PI controller by using a metaheuristic algorithm called Particle Swarm Optimization [7].

In spite of the advances already made in multivariable controller design, there are more variants coming from computational science that may still contribute by producing substantial improvements in modern controllers. It is worthwhile exploring the full potential in metaheuristics [8] for controller design. Though profitable, metaheuristic approaches have not been exploited deeply enough in the field of control design up to now.

In particular, GAs are both exploratory and exploitative methods for solving problems that are not affordable by traditional methods [4]. Their potential is enormous since GAs have a high implicit parallelism, making them numerically very efficient. On-line control provides time-consuming scenarios for industrial problems. Besides, parallel algorithms usually run a different search model with respect to sequential ones [8]. Therefore, apart from diminishing computing times, the solution quality tends to be better.

2. PROBLEM STATEMENT

Rodriguez et al. [9] developed a software package that is helpful when trying to find the optimum pipeline network for the transportation of natural gasoline. This software pinpoints the place where it is convenient to locate a storage tank that should serve to guarantee a steady feed flowrate to the processing plant downstream. Keeping in mind the practical implementation of suitable controllers, an ad hoc strategy is proposed in this paper.

The system under control is a storage tank (Fig. 1), where liquid hydrocarbons are collected by means of pipelines coming from various sources with their own inherent dynamics that was contemplated in our design. In particular, we have addressed the transport

![Figure 1: Configuration of the system under control](image-url)
of natural gasoline to be extracted from the gas fields located in the province of Santa Cruz (Argentina). From the storage tank, the natural gasoline should be sent with a steady flow to a processing plant. This storage tank is a MIMO (Multiple-Input-Multiple-Output) system. It consists of three inputs \( (F_0, F_1, F_2) \), which are the main flowrates originated from the gas fields known as Pico Truncado, Cañadón León and Cañadón Seco, and two control variables associated to the tank: liquid level (\( h \)) and outflow (\( F_3 \)). Unlike most of the typical multivariable control designs, the distinctive feature of this unconventional MIMO system is that it has three inputs, instead of merely two.

Fig. 2 shows a scheme of the proposed control strategy. Its design consists of two stages. In the first stage [10] two SISO (Simple-Input-Simple-Output) systems were considered, but the interaction between both of them was not contemplated. In this paper the second stage was developed by taking the interactions into account with MIMO controllers, the parameters also being optimized by GAs. As to the dynamics, the production from the main reservoirs (Cañadón Seco and Cañadón León) was considered. Besides, Pico Truncado was regarded as a complementary flow (\( F_2 \)) to the eventual decline of the other fields (\( F_0, F_1 \)). In view of this scenario, the flows entering the storage system are controlled with a feedforward loop. In this way, a lack of supply to the storage tank can be avoided. Therefore, the use of GAs is here proposed to design and optimize a feedback-feedforward integrated system to control the volume of natural gasoline at the storage tank, in order to warrant the supply to the processing plant.

![Block diagram of the proposed controllers.](image)

**Figure 2:** Block diagram of the proposed controllers.

### 3. METHODOLOGY

The generic framework to fulfill the desired requirements can be summarized as follows:

I. Design a SISO system

II. Introduce interaction by designing a MIMO system

III. Improve computing times via algorithmic parallelization.
As to the optimization method to be applied in search for the best parameters, we have resorted to a flexible iterative technique. Genetic Algorithms (GAs) [11] were considered because they constitute the most complete paradigm of evolutionary computation, i.e., they naturally meet all the fundamental ideas of this approach.

In Step I, PI and PID controllers were simultaneously designed via GAs, assuming that the system was decoupled [10]. Apart from taking into account the interactions, in Step II a feedforward loop was introduced in order to manage the tank volume. This method is based on adopting a transfer function like the one in Eq. 1 for each anticipative controller. In advance, we can avoid the lack of supply that may happen as a result of the depletion of any of the main wells. In this work, we report the design of the integrated control system as a bulk by means of GA optimizations.

$$G_{FF}(s) = -K_{ff} \frac{\tau_1 s + 1}{\tau_2 s + 1}$$  \hspace{1cm} (1)

The performance under different crossovers is evaluated. The parameters of controllers were optimized by means of GAs. Figure 3 shows the vector of individuals. The corresponding fitness function (Eq. 2) was formulated aiming at the minimization of the system’s error. The error $\varepsilon_{\text{PID}}$ contemplates both the PID effect and the anticipative influence of $C_{\text{FF1}}$. Similarly, the error $\varepsilon_{\text{PI}}$ contemplates both the PI effect and the anticipative influence of $C_{\text{FF2}}$. (See Fig. 1)

$$f = \frac{|\varepsilon_{\text{PID}}(t)|}{H_{SP}} + \frac{|\varepsilon_{\text{PI}}(t)|}{F_{SSP}}$$  \hspace{1cm} (2)

Step 3 is important for this design because not only heuristics that produce superior solutions were considered important, but also the computational speed was contemplated as a key factor. Then, improvements via parallelization were introduced. The adopted technique consists in using multiple processors to evaluate various individuals simultaneously.
4. COMPUTATIONAL RESULTS

The following parameters were employed for the GA optimizations: 150 generations, crossover fraction= 0.8, and 100 individuals in the population. The executions were performed on an AMD 8120 Eight-Core Processor with 3.10 GHz and 8 GB of RAM.

Experimental results for the system evolving under 6 different crossover methods, which were executed with the GA Toolbox from Matlab [12], are presented in Table 1. The reported fitness values were obtained after having simulated the system with Simulink. For the problem under study, all values are satisfactory. The Mean Best Fitness (MBF) values are close to the Best Fitness Value at Termination (BFVT). In particular, the crossover based on heuristics yielded the lowest fitness value, while the scattered crossover exhibited the best MBF.

<table>
<thead>
<tr>
<th>Crossovers</th>
<th>BFVT</th>
<th>Time (s)</th>
<th>Iterations</th>
<th>MBF</th>
<th>Average Time (s)</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scattered</td>
<td>2.0595</td>
<td>238</td>
<td>81</td>
<td>2.252</td>
<td>196</td>
<td>2.187</td>
</tr>
<tr>
<td>Two point</td>
<td>2.0291</td>
<td>239</td>
<td>78</td>
<td>2.373</td>
<td>180</td>
<td>2.234</td>
</tr>
<tr>
<td>Single point</td>
<td>2.0276</td>
<td>234</td>
<td>77</td>
<td>2.435</td>
<td>172</td>
<td>2.232</td>
</tr>
<tr>
<td>Intermediate</td>
<td>2.1601</td>
<td>267</td>
<td>79</td>
<td>2.431</td>
<td>294</td>
<td>2.520</td>
</tr>
<tr>
<td>Heuristic</td>
<td>2.0078</td>
<td>357</td>
<td>91</td>
<td>2.396</td>
<td>204</td>
<td>2.354</td>
</tr>
<tr>
<td>Arithmetic</td>
<td>2.1541</td>
<td>316</td>
<td>88</td>
<td>2.512</td>
<td>273</td>
<td>2.532</td>
</tr>
</tbody>
</table>

Performance can be assessed by means of Fig. 4, which shows the box plots associated with the following crossover methods: H= heuristic; S= scattered. In comparison with all tested crossover methods, their interquartile ranges (IQR$_H$=0.404; IQR$_S$=0.073) reveal acceptable variability in executions. Moreover, the scattered crossover exhibits lower data dispersion around the median, which is reported in Table 1.

![Figure 4: Vertical box plots illustrating Heuristic (left) and Scattered (right) crossovers](image)
Once the most convenient crossovers had been identified, parallel GAs were evolved. If computational times reported in Table 2 are compared with the sequential ones (see Table 1), a considerable reduction was achieved. Parallel techniques proved to be successful for controller design, enabling fast convergence by reducing iterations and execution times. The GAs implemented with either heuristic or scattered crossovers exhibited high efficiencies for the runs with 8 processors.

Table 2: Execution times and iteration numbers for the parallel GA programs.

<table>
<thead>
<tr>
<th>Crossover</th>
<th>BFVT Time (s)</th>
<th>Time (s)</th>
<th>Iterations</th>
<th>MBF</th>
<th>Averagen Time (s)</th>
<th>Speed-up</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heuristic</td>
<td>2.0198</td>
<td>42</td>
<td>65</td>
<td>2.382</td>
<td>37</td>
<td>5.51</td>
<td>68.91</td>
</tr>
<tr>
<td>Scattered</td>
<td>2.0236</td>
<td>37</td>
<td>74</td>
<td>2.231</td>
<td>34</td>
<td>5.76</td>
<td>72.05</td>
</tr>
</tbody>
</table>

Table 3 shows the controller parameters adjusted via parallel GA implemented with either heuristic or scattered crossovers.

Table 3: Individuals resulting from parallel runs.

<table>
<thead>
<tr>
<th>System</th>
<th>Crossovers</th>
<th>$K_p$</th>
<th>$K_i$</th>
<th>$K_d$</th>
<th>$K_p$</th>
<th>$K_i$</th>
<th>$K_f$</th>
<th>$\tau_1$</th>
<th>$\tau_2$</th>
<th>$K_f$</th>
<th>$\tau_1$</th>
<th>$\tau_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Heuristic</td>
<td>43.98</td>
<td>78.48</td>
<td>0.65</td>
<td>15.38</td>
<td>23.38</td>
<td>4.00</td>
<td>16.35</td>
<td>4.34</td>
<td>13.51</td>
<td>14.17</td>
<td>34.57</td>
</tr>
<tr>
<td>B</td>
<td>Scattered</td>
<td>53.09</td>
<td>48.72</td>
<td>1.27</td>
<td>6.56</td>
<td>57.82</td>
<td>3.73</td>
<td>31.75</td>
<td>33.36</td>
<td>0.65</td>
<td>0.47</td>
<td>3.33</td>
</tr>
</tbody>
</table>

For the developed controllers, simulations were carried out to illustrate their performance in set-point tracking. Fig. 5 shows the satisfactory behavior of system A.

Figure 5: Step responses under control system A: tank level $h$ (left); output flowrate $F_3$ (right)

5. CONCLUSIONS

A MIMO system was developed in order to find the best control parameters for a storage tank where the production coming from gas fields is temporarily stored. The analysis was focused on GA techniques with an assortment of crossover alternatives. The best performance was evidenced by the heuristic crossover, while the scattered one also reflected good quality behaviour. By using parallel processing, a significant speed-up could be noticed. The results also showed the effectiveness and reliability of the proposed approaches.
REFERENCES


PRECONDITIONED IMPLICIT RUNGE-KUTTA SCHEMES FOR UNSTEADY SIMULATIONS OF LOW MACH NUMBER COMPRESSIBLE FLOWS

LEONARDO S. DE B. ALVES

1Laboratório de Mecânica Teórica e Aplicada - LMTA, Departamento de Engenharia Mecânica - TEM, Universidade Federal Fluminense - UFF, Rua Passo da Pátria 156, Bloco E, Sala 216, Niterói, RJ 24210-240, Brazil.

Key words: Low Mach Preconditioned Density-Based Methods, Diagonally Implicit Runge-Kutta Schemes, High-Order in Time, Unconditional Numerical Linear Stability

Abstract. An extensive literature review of low Mach Preconditioned Density-Based Methods reveals that physical-time marching is almost always performed with a second-order BDF scheme. As far as the author is aware, the second-order Crank-Nicolson scheme has been the only notable but very rare exception. In other words, these methods have always employed some form of multi-step scheme in physical-time. Since these schemes lose unconditional linear stability when their order is increased above two, simulations with this method have been essentially restricted to second-order accurate in time. The present paper demonstrates for the first time how the traditional dual-time stepping procedure naturally leads to the low Mach preconditioning of the intermediate stage variables within implicit Runge-Kutta schemes. Such an extension allows the use of high-order multi-stage schemes by this method. Second, third and fourth-order diagonally implicit Runge-Kutta schemes are employed under this new methodology to simulate vortex formation in absolutely unstable planar mixing-layers, providing strong evidence of its capabilities. Results indicate that, as it should be expected, high-order multi-stage schemes with strong linear numerical stability outperform traditional lower order multi-step schemes when stringent tolerance requirements are imposed.

1 INTRODUCTION

The ever increasing computer power reduced total calculation times and memory costs to such an extent that numerical simulations of unsteady three-dimensional equations are now feasible. Selecting a time-marching scheme for unsteady simulations is a compromise between efficiency and accuracy. Furthermore, both are controlled by stiffness to a great extent, which is caused by disparities in characteristic time and/or length scales. One of the major difficulties faced by researchers performing simulations of very low speed
compressible flows is the stiffness due to large differences between convective and acoustic velocities. As a result, time-steps dictated by numerical stability are much smaller than required by accuracy considerations, drastically increasing computer time of conditionally stable time-marching schemes. In other words, stiffness decreases efficiency. Realistic computer times are only possible in such cases with numerical schemes that are at least A-stable, a mathematical property defined as unconditional stability when solving a linear and homogeneous standard test problem. An additional difficulty associated with low Mach number flows is the fact that pressure is no longer a thermodynamic state variable, but only a Lagrange multiplier. Hence, one must solve for it directly, instead of recovering it from an equation of state, to avoid significant round-off error propagation. In other words, stiffness decreases accuracy. Arguably the two main approaches advocated to deal with these difficulties and simulate low Mach number flows are known as pressure-based and density-based methods [1].

Preconditioned density-based methods are the focus of this study. Their origins trace back to the artificial compressibility for incompressible flows [2]. This idea was extended towards compressible flows twenty years later [3], allowing traditional compressible solvers to simulate low Mach number flows as well. An additional important contribution was the separation of static (thermodynamic) and dynamic (hydrodynamic) pressures, which allowed much lower Mach numbers to be reached [4]. These methods were optimized to allow simulations with high and low Reynolds number flows [5], reacting flows with non-equilibrium chemistry [6], highly stretched meshes [7], implicit upwind solvers [8], different variable sets [9], reduced eigenvector orthogonality losses [10] and improved energy conservation accuracy at very low Mach numbers [11, 12]. Extensive literature reviews can be found elsewhere [13, 14]. Nevertheless, preconditioning alters the original physical-time evolution, although a correct steady-state is always recovered. Hence, unsteady simulations require dual time stepping [15], a technique originally introduced with artificial compressibility [16]. However, low Mach number preconditioned density-based methods are usually limited to second-order accuracy in physical-time [17]. In fact, the author has not been able to find a low Mach preconditioned density-based method in the peer-reviewed journal literature that has an accuracy order higher than two in physical-time [18]. This restriction is due to the fact that these methods always employ multi-step schemes for an accurate temporal resolution, but there are no A-stable multi-step schemes with accuracy orders greater than two [19]. Only a few higher order versions are available, but they are all conditionally stable [20].

Conditional stability in higher order implicit multi-step schemes led several researchers towards their multi-stage counterparts to minimize the CFL restrictions imposed by stiffness [21, 22]. They are known as implicit Runge-Kutta (IRK) methods [23]. High accuracy-order with strong numerical stability is achieved through an increase in the number of intermediate stages used between physical-time steps. These methods are self-starting as well, and easy to implement with variable time-step sizes, which is not the case for multi-step schemes.
Nevertheless, not even the stronger numerical stability of IRK methods is able to control the stiffness found in very low Mach number flows, where $M \ll 1$, evidenced by the fact that lower bounds in previously mentioned studies were limited to $M \sim 0.05$. The main goal of the present study is to demonstrate how the well-known low Mach preconditioning techniques can be employed with these multi-stage schemes as well. This is achieved in such a way that no modifications whatsoever are required of the preconditioning matrix and preconditioned artificial dissipation for spatial resolution, allowing a straightforward use of all scientific developments related to preconditioned density-based methods achieved over the past twenty-five years. The present work provides the framework for low Mach preconditioned density-based methods to be employed with multi-stage schemes, allowing high-order unsteady simulations with strong numerical stability of very low Mach number compressible flows. Absolutely unstable planar mixing-layers are simulated using this novel methodology, providing strong evidence for the improved efficiency and accuracy of this novel approach when compared to non-preconditioned multi-stage methods and low Mach preconditioned multi-step schemes.

2 GOVERNING EQUATIONS

Consider the 2D compressible Euler equations from fluid dynamics, written as

$$\frac{\partial q}{\partial t} = f(q)$$

where $f(q) = -\left(\frac{\partial e_i}{\partial \xi} + \frac{\partial f_i}{\partial \eta}\right)$, (1)

where $t$ is the physical-time independent variable and $\xi$ and $\eta$ are the independent spatial variables in computational space. Their differentials are related to their counterparts in physical space, $x$ and $y$, through the Jacobian determinant $J$. Furthermore, the generalized dependent variable $q$ and inviscid fluxes $e_i$ and $f_i$ in the steady-state residue $f(q)$ are related to their counterparts in physical space through

$$q = J Q, \quad e_i = J E_i \frac{\partial \xi}{\partial x} + J F_i \frac{\partial \xi}{\partial y} \quad \text{and} \quad f_i = J E_i \frac{\partial \eta}{\partial x} + J F_i \frac{\partial \eta}{\partial y},$$

which, in turn, are defined in conservative form as

$$Q = \{ \rho, \rho u, \rho v, \rho E \}$$

$$E_i = \{ \rho u, \rho u^2 + P, \rho u v, (\rho E + P)u \} \quad \text{and}$$

$$F_i = \{ \rho v, \rho u v, \rho v^2 + P, (\rho E + P)v \}$$

where $\rho$ stands for density, $u$ for stream wise velocity, $v$ for cross stream velocity, $E = e + (u^2 + v^2)/2$ for total internal energy per unit mass, $e$ for thermal internal energy per unit mass and $P$ as pressure.
In order to generate discrete solutions from these conservation equations at low Mach numbers, two different techniques must be employed. First, it is important to solve for pressure instead of density, since the latter becomes independent of the former at low Mach numbers. Furthermore, pressure should be decomposed into \( P = P_T + P_H \), which is the sum of its hydrodynamic \( P_H \sim \rho u^2 \) and thermodynamic \( P_T \sim \rho c_s^2 \) contributions, since the latter is essentially constant and orders of magnitude higher than the former when \( M \ll 1 \). Such a step is essential for controlling pressure roundoff error propagation in the momentum conservation equations, where \( P \) must be replaced by \( P_H \). It should be noted that \( P_T \) is constant in space, but can be allowed to vary in time as is often the case in fully bounded flows. Nevertheless, \( P_T \) is assumed constant here. Equation (1) then becomes

\[
T \frac{\partial \hat{q}}{\partial t} = f(q) ,
\]

which now uses the primitive dependent variable vectors

\[
\hat{q} = J \hat{Q} \quad \text{and} \quad \hat{Q} = \{ P_H, u, v, T \} ,
\]

and the conservative to primitive variable Jacobian \( T \). Although pressure and velocities are often present in \( \hat{Q} \), temperature can be replaced by enthalpy or entropy, among others. Density and enthalpy dependencies on pressure and temperature are determined from an equation of state.

Despite these modifications, Eq. (6) becomes increasingly stiff as the Mach number is decreased below \( M \sim 0.1 \). Simulations with smaller Mach numbers require the second technique, i.e., low Mach preconditioning. It rescales the eigenvalues of (6), forcing them to have the same order of magnitude even when \( M \ll 1 \). This is achieved by replacing conservative to primitive variable Jacobian \( T \) by a preconditioning matrix \( \Gamma \), leading to

\[
\Gamma \frac{\partial \hat{q}}{\partial \tau} = f(q) ,
\]

where \( \Gamma \) can be taken from any standard references, such as [14].

However, preconditioning alters the original physical-time \( t \), creating a pseudo-time \( \tau \) in which the flow now evolves. Nevertheless, it accelerates convergence of the original system, in the form of either Eq. (1) or (6), towards steady-state. Upon reaching this limit, the same solution of \( f(q) \simeq 0 \) is recovered from preconditioned system (8) as well. Physical-time accuracy is still possible for preconditioned density-based methods, but only if the original physical-time derivative is re-introduced into (8), yielding

\[
\Gamma \frac{\partial \hat{q}}{\partial \tau} + \frac{\partial \hat{q}}{\partial t} = f(q) ,
\]

a procedure known as dual time stepping, or DTS. Whenever pseudo-time steady state is reached within each physical-time step, the original governing Eq. (1) is recovered.
In order to implement multi-stage schemes for physical-time marching in low Mach preconditioned density-based methods, one must first treat the preconditioned pseudo-time derivative in Eq. (9) as a source term, leading to

$$\frac{\partial q}{\partial t} = \bar{f}(q) = f(q) - \Gamma \frac{\partial \hat{q}}{\partial \tau} ,$$

(10)

where $\bar{f}(q)$ is the pseudo-unsteady residue. Then, the above equation is marched in physical-time by a general IRK method using

$$q^{n+1} = q^n + \Delta t \sum_{i=1}^{k} \omega_i \bar{f}(k_i) ,$$

(11)

where the $k$ intermediate stage variable vectors $k_i$ are obtained from

$$k_i = q^n + \Delta t \sum_{j=1}^{k} \beta_{i,j} \bar{f}(t_n + \delta_i \Delta t , k_j)$$

$$= q^n + \Delta t \left( \sum_{j=1}^{i-1} \alpha_{i,j} \bar{f}(k_j) + \beta_i \bar{f}(k_i) + \sum_{j=i+1}^{k} \gamma_{i,j} \bar{f}(k_j) \right) \text{ for } 1 \leq i \leq k ,$$

(12)

with the term $t_n + \delta_i \Delta t$ being removed for clarity purposes. Coefficients $\alpha_{i,j}$, $\beta_i$ and $\gamma_{i,j}$ – which are the lower, diagonal and upper coefficients of $\beta_{i,j}$, respectively – as well as $\delta_i$ and $\omega_j$ play their well known roles in Runge-Kutta schemes. $\delta_i$ is omitted whenever possible to simplify the notation. Explicit RK (ERK) methods are obtained with $\beta_i = \gamma_{i,j} = 0$, diagonally implicit RK (DIRK) methods with $\gamma_{i,j} = 0$ and fully implicit RK (FIRK) methods with all coefficients being generally nonzero. Most IRK methods are subsets of these three groups. Next, both steady-state residue and preconditioned pseudo-time derivative inside the pseudo-unsteady residue are separated in Eq. (12) to yield

$$\sum_{j=1}^{k} \beta_{i,j} \Gamma_j \frac{\partial \hat{k}_j}{\partial \tau} = \frac{q^n - k_i}{\Delta t} + \sum_{j=1}^{k} \beta_{i,j} f(k_j)$$

(13)

for $1 \leq i \leq k$. Furthermore, $\hat{k}_i$ are the intermediate variable vectors in primitive form. They are related to their conservative counterparts $k_i$ in the same way $\hat{q}$ is related to $q$. Nevertheless, this equation is written in such a way that its numerical solution is obtained very inefficiently. This is caused by the existence of multiple pseudo-time derivatives in each intermediate stage. One may note, however, that the summations on both sides have the same coefficients. Hence, this equation can be re-written as

$$\Gamma_i \frac{\partial \hat{k}_i}{\partial \tau} = \bar{f}(k_i) = f(k_i) + \sum_{j=1}^{k} \beta_{i,j}^{-1} \left( \frac{q^n - k_j}{\Delta t} \right) \text{ for } 1 \leq i \leq k ,$$

(14)
where $\beta_{i,j}^{-1}$ represents the coefficients of the inverse of the matrix whose coefficients are $\beta_{i,j}$ and, hence, can be determined \textit{a priori} for arbitrary $k$ values. $\tilde{f}(k_i)$ is now a pseudo steady-state residue for the intermediate stages. Once pseudo-time steady-state is reached for all intermediate stage equations within each physical-time step, Eq. (14) converges to its original non-preconditioned version $\tilde{f}(k_i) \simeq 0$. Equation (11) can be then utilized for the physical-time update of Eq. (1) with $\tilde{f}(k_i)$ replaced by $f(k_i)$, since they become approximately equal when pseudo-time steady-state is reached.

The explicit coupling terms that appear in Eq. (14) have a peculiar characteristic. They have the form of a first-order first derivative approximation. It is also important to note that this explicit coupling in DIRK schemes retains the same triangular nature of the implicit coupling from their original versions in Eq. (12). In fact, all preconditioned implicit RK (PIRK) methods maintain the original nature of their respective implicit couplings, although in an explicit manner. A less obvious characteristic of this approach is its inability to generate a preconditioned ERK (PERK) method for physical-time marching. This happens because the pseudo-time derivative is introduced through $f(q)$, which does not appear when calculating $k_i$ for explicit ($\beta_i = 0$) RK methods.

Equation (14) is marched in pseudo-time with the implicit Euler method

$$\Gamma_i^p \frac{\hat{k}^{p+1}_i - \hat{k}^p_i}{\Delta \tau} \simeq \tilde{f}(k^{p+1}_i) \quad \text{for} \quad 1 \leq i \leq k,$$

which, upon linearization, becomes

$$\left\{ \frac{\Gamma_i}{\Delta \tau} - \frac{\partial \tilde{f}_i}{\partial \hat{k}_i} \right\}^p \Delta \hat{k}_i \simeq \tilde{f}(k^p_i) \quad \text{for} \quad 1 \leq i \leq k,$$

where $\hat{k}^{p+1}_i = \hat{k}^p_i + \Delta \hat{k}_i$ and $\partial \tilde{f}_i / \partial \hat{k}_i$ is a general Jacobian equivalent to $\partial f / \partial q$. Combining Eqs. (14) and (16) yields

$$\left\{ \frac{\Gamma_i}{\Delta \tau} - \frac{\partial \tilde{f}_i}{\partial \hat{k}_i} \right\}^p \Delta \hat{k}_i + \sum_{j=1}^{k} \beta_{i,j}^{-1} T_j^p \Delta \hat{k}_j \simeq f(k^p_i) + \sum_{j=1}^{k} \beta_{i,j}^{-1} \left( \frac{q^n - k_j^p}{\Delta t} \right),$$

for $1 \leq i \leq k$, where $T_j = \partial k_j / \partial \hat{k}_j$ is now the conservative to primitive intermediate stage variable vector Jacobian and $\partial \tilde{f}_i / \partial \hat{k}_i$ contains the remaining Jacobians, just as $\partial f / \partial q$ does. The development of this equation considered the general case of a fully implicit RK scheme. At any given stage in diagonally implicit RK schemes, intermediate stage variables from previous and posterior stages are already known and don’t exist, respectively. In such cases, the equation above can be re-written as

$$\left\{ \frac{\Gamma_i}{\Delta \tau} + \beta_{i,i}^{-1} T_i \frac{\partial \tilde{f}_i}{\partial \hat{k}_i} \right\}^p \Delta \hat{k}_i \simeq f(k^p_i) + \sum_{j=1}^{i-1} \beta_{i,j}^{-1} \left( \frac{q^n - k_j^{p+1}}{\Delta t} \right) + \beta_{i,i}^{-1} \frac{q^n - k_i^p}{\Delta t},$$

for $1 \leq i \leq k$.
for $1 \leq i \leq k$. Second, third and fourth-order accurate SDIRK schemes that are either A-stable or strongly S-stable were chosen for the present study. The latter property is equivalent to L-stability, but when applied to a linear but nonhomogeneous standard test problem. FIRK schemes were also tested but are not shown because they are not competitive with or without preconditioning. Details about the form taken by Eq. (18) for each IRK scheme and its respective preconditioned version are given elsewhere [24].

The main advantage of using either Eq. (17) or (18), obtained from this novel PIRK methodology presented here, is their consistency with Eq. (9). It follows the traditional DTS approach employed by preconditioned density-based methods to simulate unsteady low Mach number flows. Hence, all the low Mach preconditioning techniques employed over the past two and a half decades can now be directly used with multi-stage schemes as well without any modifications whatsoever to either the preconditioning matrix or the preconditioned spatial discretization.

4 RESULTS AND DISCUSSION

In every test case reported next, the number of spatial grid points was made large enough in order to guarantee that numerical resolution is dominated by temporal instead of spatial errors. High-order spatial resolution for the inviscid fluxes (4) and (5) in Eq. (1) is achieved with a fifth-order version [25] of the well-known flux-difference method with a preconditioned artificial dissipation matrix [8]. Furthermore, there is periodicity in the streamwise direction. Spatial periodicity eliminates the need to use boundary closure schemes and, hence, avoids an additional error source that would otherwise contaminate the numerical solution. Periodic boundary conditions are exactly enforced in the explicit residue of Eq. (18), but are approximated in its implicit operator to avoid a cyclic matrix structure. Cross stream boundary conditions employ a zero derivative approximation for all variables but pressure, which is fixed. No buffer zones are necessary in these side boundaries because initial conditions are constructed to satisfy the steady governing equations, essentially eliminating initial condition noise. However, grid stretching was employed in this direction away from the physically relevant domain region to damp any perturbations that reach the cross stream boundaries. Additionally, the scheme used for the inviscid flux was altered near these boundaries to reduce its order to four, enhancing numerical stability of the overall method.

The planar mixing-layer initial condition is first modelled with a well known profile from the literature [26] and perturbed with approximate expressions [27]. Since they are not solutions of the governing equations, this perturbed initial condition generates numerical error, which was damped using Physical-Time Damping or PTD [28]. When this numerical error became negligible, a solution was extracted and used as a highly accurate initial condition perturbed with the desired linear modes.

Nine different wave number excitation cases are analysed here. Frequencies (left) and temporal growth rates (right) extracted from these simulations are shown in Fig. 1, which also shows the behavior predicted by linear stability analysis (solid lines). An excellent
agreement can be observed for the frequency data whereas non-negligible, although still relatively small, differences can be observed for the temporal growth rates. It is important to note that such a higher error for the growth rate, compared to its respective frequency, is observed even in the linear stability analysis calculations. These errors are sensitive to the spatial location chosen for the extraction of the temporal data points, but they always stay approximately within the minimum and maximum values presented in the legend of this figure. Nevertheless, temporal growth rate relative errors do appear to systematically increase with wave number, which is consistent with the dissipative error expected of biased upwind schemes such as the one employed for the present spatial discretization. Despite these issues, the present results are well within the accuracy bounds found in the cited literature and, hence, provide sufficient validation for the novel methodology applied to the planar mixing-layer. Furthermore, all the other multi-step and multi-stage schemes tested provide similar results.

After code validation, an accuracy-order verification of the different schemes studied in this paper is now performed. In order to do so in a concise manner, these results are presented only for the $k_R^* = 0.4$ case. Absolute errors are estimated using the numerical solution obtained with twice as many points per period. Figure 2 shows dimensionless frequency and temporal growth rate absolute errors, which are the real and imaginary parts of $\Delta \omega / \omega_0$, respectively, versus dimensionless physical-time step for all four second-order-methods analysed, where dashed lines indicate the theoretical second, third and fourth-order slopes. Oscillations are likely caused by the order loss phenomenon found in high-order multi-stage schemes, likely triggered by additional errors due to the fact that error measurements are not taken directly from the simulated data, but are instead based on frequencies and temporal growth rates calculated from the simulated data. Whenever temporal accuracy-order increases, error measurements decrease towards their respective minimal standard deviations found in the nonlinear fitting procedure used to estimate
frequencies and growth rates. This problem could be avoided by employing larger data samples, which would reduce the minimal standard deviations. However, such a procedure would require either the use of initial perturbations with even smaller amplitudes or the simulation of linearized equations instead. The former would increase much further the CPU time required for the present study whereas the latter is beyond the scope of the present study. Neither approach is necessary though, given the reasonable qualitative agreement between theoretical and numerical slopes observed in Fig. 2.

Finally, an efficiency analysis for the two-dimensional test case is shown in Fig. 3. Efficiency is defined as the total iteration count \( \tilde{m} \cdot N_T \) required to generate a solution for \( \Delta \omega / \omega_0 \) with a prescribed tolerance, be it frequency or temporal growth rate, since the computer time per pseudo-time iteration is essentially the same for all multi-step and multi-stage schemes studied here. In general, Fig. 3 indicates that higher-order and/or stronger numerical stability lead to more efficient methods. When focusing on frequency data, shown in Fig. 3 (left), PCN2-AS is the most efficient choice when errors higher than 3% can be tolerated. Otherwise, PSDIRK2-SSS is the optimal choice, but only when errors higher than 0.6% can be tolerated. If smaller errors are required, PSDIRK3-SSS is the best available choice among the schemes analyzed. Similar conclusions can be reached
from temporal growth rate data as well, which is shown in Fig. 3 (right). PCN2-AS is still the most efficient choice, but only when tolerances higher than 9% are imposed. When smaller errors are required, PSDIRK2-SSS is the preferred choice.

5 CONCLUSIONS

Recently, temporal integration in high Mach number compressible flow simulations has achieved higher accuracy orders and/or stronger numerical stability by using multi-stage schemes. This study shows that the same can be done for low Mach preconditioned density-based methods. Furthermore, such a feat is achieved using the same low Mach preconditioned dual-time-stepping procedure traditionally employed by these methods, without introducing any additional approximations. This means that all low Mach preconditioning matrixes and all low Mach preconditioned artificial dissipation schemes for spatial resolution developed over the past three decades can now be used with multi-stage physical-time integration schemes, requiring no modifications or adjustments whatsoever. In doing so, the large field of research devoted to Runge-Kutta schemes is now open to low Mach preconditioned density-based methods as well.

ACKNOWLEDGEMENTS

The author would like to express his gratitude for the financial support from FAPERJ through grant E-26/103.254/2011 and CNPq through grant 481072/2012-8.

REFERENCES


NUMERICAL INVESTIGATION OF TURBULENT CHANNEL FLOWS USING A SEMI-SPECTRAL CODE

L.A.C.A. SCHIAVO\textsuperscript{1,4}, A.B. JESUS\textsuperscript{2}, J.L.F. AZEVEDO\textsuperscript{3}, W.R. WOLF\textsuperscript{4}

\textsuperscript{1,4}Universidade Estadual de Campinas, UNICAMP
Rua Mendeleyev, 200, 13083-860, Campinas, SP, Brazil
\textsuperscript{1}augustoschiavo@gmail.com \textsuperscript{4}wolf@fem.unicamp.br

\textsuperscript{2}Instituto Tecnológico de Aeronáutica, DCTA/ITA
Praça Mal. Eduardo Gomes, 50, 12228-900, São José dos Campos, SP, Brazil
antoniobjota@gmail.com

\textsuperscript{3}Instituto de Aeronáutica e Espaço, DCTA/IAE/ALA
Praça Mal. Eduardo Gomes, 50, 12228-904, São José dos Campos, SP, Brazil
joaoluiz.azevedo@gmail.com

Key words: Turbulent flow, Channel flow, LES, Adverse pressure gradient region

Abstract. In the present work, wall-resolved Large Eddy Simulations (LES) of incompressible turbulent flows are presented for a convergent-divergent channel with adverse pressure gradients. The paper presents comparative results for different Reynolds numbers, studying wall friction coefficients and turbulent kinetic energy budgets from LES computations. The turbulent kinetic energy budgets are investigated at the separation and at re-attachment points. The CFD code used in this work is based on the incompressible Navier-Stokes equations, solved for structured grids with a semi-spectral methodology. The simulations performed are wall-resolved and the WALE subgrid scale model is used for turbulence closure. The analyses of the turbulent kinetic energy budgets find specific distribution patterns, respectively, at separation and re-attachment points.

1 INTRODUCTION

Turbulence is observed in several flows present in nature and in industrial applications. Turbulent flows are characterized by a large disparity of spatial and temporal scales. In order to accurately resolve these scales, direct numerical simulation (DNS) and large eddy simulation (LES) have been used to study the physical phenomena associated with transition and turbulence. While DNS resolves all ranges of scales in a turbulent flow, in LES, the larger scales, which are mostly affected by the topology of the flow, are effectively resolved while the high wavenumber turbulent scales are modeled by a subgrid
scale model. One should mention that these smaller scales are less energetic and their statistics have a more universal character.

In industrial applications, the accurate prediction of separation and re-attachment points in turbulent flows is an important factor in the design of aircraft, launch vehicles and turbo-machinery. Detachment regions generally produce a significant reduction of the performance of turbo-machinery and aircraft. Adverse pressure gradients (APG) change the shear stress distribution, and this can often change the dynamics of the flow and be responsible for flow separation. Currently, the application of CFD tools which resolve the Reynolds-averaged Navier-Stokes (RANS) equations is a common practice in industry. However, RANS turbulence models are still not capable of adequately predicting the flow characteristics in APG regions [1, 2] in spite of recent improvements to such models. Two-equation models tend to either under-predict or over-predict flow separation. Reynolds stress transport models show improved results, especially in the pressure distribution, yet none of the RANS models proved to be fully accurate in skin friction calculations. The inaccuracies of RANS models in accurately predicting the flow physics along regions with APG have been associated with the inability of such models of correctly capturing flow separation and re-attachment locations [2, 3]. On the other hand, recent studies [1] suggest that the behavior of the skin friction coefficient in APG regions of smooth bumps does not depend on flow separation. In other words, the behavior of attached and mildly separated flows is very similar, indicating that the prevailing physical mechanism is mostly associated to the adverse pressure gradient condition.

In this work, wall-resolved LES is applied to study incompressible flows along channels with contractions at low and moderate Reynolds numbers. The effects of adverse pressure gradients are evaluated through the analysis of distribution of friction coefficient and budgets of turbulent kinetic energy including the role of production, transport and dissipation of turbulence. Similar studies were performed by [4] and [5] for low Reynolds numbers and the present LES results show excellent agreement compared to DNS solutions available in the literature.

2 BUMP GEOMETRY

The bump profile under study is presented in Fig. 1. It combines an initial favorable pressure gradient region in the upstream portion of the bump with an APG region in the back of it. This bump profile was studied both in wind-tunnel experiments, at high Reynolds numbers, and by DNS calculations [4, 6], at low Reynolds numbers, $Re_t = 180, 395$ and $617$, during the Wallturb European project [7]. For all calculations performed here, as well as in the DNS computations, the bump is considered as installed in a channel with $2.0$ in height and $4\pi$ in length.
3 NUMERICAL FORMULATION

LES computations are performed using the MFLOPS3D code, developed at Laboratoire de Mécanique de Lille (LML). This is a semi-spectral code, developed for the study of boundary layer and channel flows including geometrical bumps. The code has been used for performing laminar flow instability computations [8], DNS of turbulent channel flows around two-dimensional bumps [4, 6] and LES of similar flow configurations [9]. The code, when running in LES mode, solves the filtered incompressible Navier-Stokes equations, which can be written as

\[
\begin{align*}
\text{Mass conservation:} & \quad \nabla \cdot \vec{u} = 0 , \\
\text{Momentum:} & \quad \frac{\partial \vec{u}}{\partial t} + (\vec{u} \cdot \nabla) \vec{u} = -\nabla p + \frac{1}{Re} \nabla^2 \vec{u} - \nabla \cdot \tau^{sf}_{ss} , \\
\text{Poisson for pressure:} & \quad \nabla^2 p = -\nabla \cdot (\vec{u} \cdot \nabla) \vec{u} - \nabla^2 \tau^{sf}_{ss} .
\end{align*}
\]  

Equation (1) represents mass and momentum conservation as well as the Poisson equation. In Eq. (1), \( \vec{u} \) represents the velocity vector, \( p \) is the pressure, \( t \) is the time and \( \tau^{sf}_{ss} \) is the subgrid scale tensor. The bars, \( \bar{\,} \), indicate filtered variables, and the subgrid scale tensor in indicial notation is defined as

\[
\tau^{sf}_{ij} = \bar{u}_i \bar{u}_j - \bar{\bar{u}}_i \bar{\bar{u}}_j .
\] 

In the present paper, subgrid scale terms are evaluated using the WALE model [10], which is very adequate for wall bounded flow computations as it was conceived to recover the correct eddy-viscosity near wall scaling without the need for an explicit damping.

The transformed Navier-Stokes equations are discretized using an 8th-order centered finite difference scheme in the streamwise direction for the first derivatives. The viscous terms are discretized using a 4th-order centered finite difference scheme. Chebyshev polynomials, collocated in Chebyshev-Lobatto points [11], are employed in the wall normal direction. A Fourier transform is performed in the spanwise direction, which is assumed to be periodic, using the 3/2 rule to remove aliasing error of the discrete Fourier transform [11]. Time integration is performed using an implicit 2nd-order backward Euler method for the terms of the Laplacian operator, while an explicit 2nd-order Adams-Bashforth
method is used for all other terms, including the sub-filter stresses. Pressure-velocity coupling is achieved by a fractional-step method which performs an iterative process. In this process, the solution of the momentum equations yield an intermediate velocity field, whereas the solution of the pressure Poisson equation determines an intermediate pressure \[12\]. Afterwards, iterations based on the continuity equation are used in order to obtain a pressure correction that produces a divergent-free velocity field. Computations are performed in Fourier space, and each Fourier mode is solved independently using parallel computations. The nonlinear and sub-filter terms are computed in physical space. The computational process is parallelized by mesh partition and the Message Passing Interface (MPI) protocol is used to communicate between processors.

Further details of the transformed equations and numerical methodology can be found in previous work which uses the MFLOPS3D code \[4\]. Boundary conditions for the inlet flow are obtained from a precursory periodic channel flow computation. No-slip conditions are imposed at the top and bottom walls, and the spanwise direction is assumed periodic with a \(\pi\) width. The outflow boundary condition is described as

\[
\frac{\partial \vec{u}}{\partial t} + U_c \frac{\partial \vec{u}}{\partial x} = 0 .
\] (3)

Therefore, the channel exit plane is treated as a convective boundary with uniform velocity \(U_c\) which transports the velocity vector field.

4 RESULTS

In the present work, wall-resolved LES computations at \(Re_\tau = 617\) and 950 are performed. Periodic channel flows at the same Reynolds numbers are also computed and, after statistical convergence, the velocity vectors at a plane normal to the streamwise direction, at the exit of the channel, are stored for two channel flow-through periods, for each case. Such velocity history is imposed as entrance condition at the inlet for the channel with contraction calculations. Computations at \(Re_\tau = 617\) use \(512 \times 129\) nodes in the streamwise and normal directions, respectively, and \(128\) Fourier modes in the spanwise direction. Simulations at \(Re_\tau = 950\) use two refinement levels. The first one corresponds to a coarse mesh employing \(448 \times 97\) grid points in the streamwise and normal directions, and \(128\) modes in the spanwise direction. A second finer level employs \(1024 \times 193\) grid points in the streamwise and normal directions, and \(256\) Fourier modes in the spanwise direction. Table 1 shows the mesh configuration and the spatial resolution in wall units at the top of the bump. Kuban et al. \[9\] performed both DNS and LES calculations, with different mesh refinements, in the same bump configuration at \(Re_\tau = 617\). Their paper showed that assuming \(\Delta x^+ \sim 45, \Delta y^+ \sim 0.25\) and \(\Delta z^+ \sim 30\) at the top of the bump provides sufficient resolution in LES calculations to allow the computation of flow statistics in good agreement with DNS results for the same configuration. Table 1 shows that the mesh refinement levels used in the present work are similar
Table 1: Mesh configurations for LES test cases.

<table>
<thead>
<tr>
<th>Reτ</th>
<th>Nx</th>
<th>Ny</th>
<th>Nz</th>
<th>Statistics Domain</th>
<th>Δx⁺</th>
<th>Δy⁺</th>
<th>Δz⁺</th>
</tr>
</thead>
<tbody>
<tr>
<td>617</td>
<td>512</td>
<td>129</td>
<td>128</td>
<td>$4\pi \times 2 \times \pi$</td>
<td>23.0</td>
<td>0.17</td>
<td>21.2</td>
</tr>
<tr>
<td>950</td>
<td>448</td>
<td>97</td>
<td>128</td>
<td>$4\pi \times 2 \times \pi$</td>
<td>24.0</td>
<td>0.45</td>
<td>31.0</td>
</tr>
<tr>
<td>950</td>
<td>1024</td>
<td>193</td>
<td>256</td>
<td>$4\pi \times 2 \times \pi$</td>
<td>16.3</td>
<td>0.11</td>
<td>15.0</td>
</tr>
</tbody>
</table>

...to those from [9]. In addition, current LES results are compared with DNS data, when available, and with other results from LES with higher resolution. In general, very good
agreement is observed in such comparisons. Therefore, it is believed that the current LES meshes are adequate to resolve the physics of interest and, therefore, to compute the flow statistics in the range of Reynolds numbers analyzed in this study.

Figure 4 shows the mean field of the $u$ velocity component and streamlines from LES results at $Re_τ = 617$. Figure 5 presents the skin friction coefficient distributions at the bottom wall for the Reynolds numbers under study. Similarly, one can see that LES results at $Re_τ = 617$ show good comparison with those obtained from DNS performed by [6]. At this Reynolds number, LES calculations show a small separation region at the top of the bump, followed by flow re-attachment still over the bump. Such results, again, are completely consistent with the DNS calculations. In both DNS and LES results, it can be seen that, after the initial decrease in the skin friction coefficient as the flow goes over the top of the bump, a strong recovery of $C_f$ is observed downstream of the separated flow region. The results for $Re_τ = 950$ demonstrate that mesh convergence has been achieved for this case. Moreover, $C_f$ curves indicate a smaller separation region at $Re_τ = 950$, when compared to the lower $Re$ case. The change in the separation and re-attachment points, for different Reynolds numbers, are also clearly seen in the figure.

The Reynolds stress transport equation is given by

$$\frac{\partial \langle u_i' u_j' \rangle}{\partial t} + \langle u_k \rangle \frac{\partial \langle u_i' u_j' \rangle}{\partial x_k} = P_{ij} + T_{ij} + D_{ij} + D_{\rho ij} + \Phi_{ij} - \epsilon_{ij}.$$  \hspace{1cm} (4)

This equation indicates the balance of advection, production, turbulent transport, viscous diffusion, pressure diffusion, pressure strain and dissipation for the terms of the Reynolds
stress tensor. The terms in the right-hand side are defined as:

\[
P_{ij} = -\langle u_j' u_k' \rangle \frac{\partial \langle u_i \rangle}{\partial x_k} - \langle u_i' u_k' \rangle \frac{\partial \langle u_j \rangle}{\partial x_k}, \quad (5)
\]

\[
T_{ij} = -\frac{\partial \langle u_i' u_j' u_k' \rangle}{\partial x_k}, \quad (6)
\]

\[
D_{ij} = \nu \frac{\partial^2 \langle u_i' u_j' \rangle}{\partial x_k^2}, \quad (7)
\]

\[
D_{pij} = -\frac{1}{\rho} \left( \frac{\partial \langle u_i' p' \rangle}{\partial x_i} + \frac{\partial \langle u_j' p' \rangle}{\partial x_j} \right), \quad (8)
\]

\[
\Phi_{ij} = \left\langle \frac{p'}{\rho} \left( \frac{\partial u_i'}{\partial x_j} + \frac{\partial u_j'}{\partial x_i} \right) \right\rangle, \quad (9)
\]

\[
\epsilon_{ij} = 2 \nu \left\langle \frac{\partial u_i'}{\partial x_k} \frac{\partial u_j'}{\partial x_k} \right\rangle, \quad (10)
\]

and the turbulent kinetic energy can be obtained by the contraction of the free indexes,

\[
k = \frac{1}{2} \langle u_i' u_i' \rangle. \quad (11)
\]
The pressure strain term vanishes in the turbulent kinetic energy balance due to the incompressibility condition, $\nabla \cdot \vec{u}' = 0$.

Detailed analyses of the turbulent kinetic energy (TKE) for the initial straight channel simulation, performed at $Re_{\tau} = 617$, is plotted against DNS data in Fig. 6. The DNS calculations were performed at $Re_{\tau} = 590$ by [13]. All balance components are normalized by $u_{\tau}^4/\nu$. The comparison shows good agreement between DNS and LES data, which, then, indicates that the present LES calculations are able to capture similar physics as the DNS results near the wall.

Figure 6: Turbulent kinetic energy budgets for flat channel: LES (grey lines) at $Re_{\tau} = 617$ and DNS (black lines) results at $Re_{\tau} = 590$.

Figure 7 shows a comparison between the turbulent kinetic energy (TKE) budgets for $Re_{\tau} = 617$ and 950 for the separation point. Here, the analyzed data are normalized by $u_{\tau_0}^4/\nu$, where $u_{\tau_0}$ is $u_\tau$ computed at the channel inlet. The data are plotted in lines normal to the wall as a function of $y^*$, where $y^*$ is defined similar to $y^+$, but computed using $u_{\tau_0}$. Figure 8 presents a similar comparison for the re-attachment point. In both figures, the solid lines represent $Re_{\tau} = 950$ data and dashed lines represent $Re_{\tau} = 617$ data. In spite of the fact that, separation points, and also re-attachment points are in different locations for each Reynolds numbers, we can observe a similar pattern in TKE budgets. In other words, the data from different Reynolds numbers actually have an interesting very good correlation at separation and re-attachment points. Furthermore, Figs. 7 and 8 also indicate that, at the separation point, there is a weakening of the viscous terms in
the balance, that is, viscous diffusion and the dissipation decrease near the wall. On the other hand, at the re-attachment point, one can observe a rapid increase of the viscous components in the budget near the wall.

Figure 7: Turbulent kinetic energy budgets for Wallturb bump geometry at the separation points: (a) LES at $Re_\tau = 617$, displaying normal lines located at $x = 5.74$. (b) LES results at $Re_\tau = 950$, displaying normal lines located at $x = 5.81$. 
Figure 8: Turbulent kinetic energy budgets for Wallturb bump geometry at the re-attachment points: (a) LES at $Re_\tau = 617$, displaying normal lines located at $x = 6.65$. (b) LES results at $Re_\tau = 950$, displaying normal lines located at $x = 6.32$.

5 CONCLUDING REMARKS

This work presents wall-resolved LES results for the skin friction coefficient along the bottom wall of a two-dimensional bump in a channel. Calculations performed at $Re_\tau =$
show a good comparison with DNS results available in the literature. Calculations at a higher Reynolds number, $Re_\tau = 950$, show $C_f$ distributions which are qualitatively very similar to those obtained at the lower Reynolds number. In particular, a sharp decrease and recovery of $C_f$ is observed in the adverse pressure gradient regions at all Reynolds numbers. Clearly, however, the extension of the separation region is smaller at $Re_\tau = 950$.

The turbulent kinetic energy budgets show a very good comparison with DNS data for a flat channel. For a channel with a contraction, the evolution of turbulent kinetic energy (TKE) budgets in the separation and re-attachment points show very similar patterns, regardless of the Reynolds numbers. Despite the fact that separation and re-attachment points are in different locations for different Reynolds numbers, the TKE budgets are very similar at the corresponding points. The results seem to suggest that there may be a standard distribution for the TKE budgets both at separation and re-attachment points, independently of the specific Reynolds number for that particular calculation. Nevertheless, more studies are necessary in order to further explore such conclusions.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the partial support for this research provided by Conselho Nacional de Desenvolvimento Científico e Tecnológico, CNPq, under the Research Grants No. 309985/2013-7, No. 400844/2014-1 and No. 443839/2014-0. This work is also supported by Fundação de Amparo à Pesquisa do Estado de São Paulo, FAPESP, under the Research Grants No. 2008/57866-1, No. 2013/03413-4, No. 2013/07375-0 and No. 2014/07623-6. The computational resources provided by Centro Nacional de Supercomputação, CESUP, of Universidade Federal do Rio Grande do Sul, and by Centro Nacional de Processamento de Alto Desempenho at UNICAMP, CENAPAD-SP, are also gratefully acknowledged.

REFERENCES


A NUMERICAL STUDY OF THE EFFECTS OF TRAILING EDGE BLUNTNESS ON AIRFOIL TONAL NOISE

Arias Ramírez W.¹, Wolf W.R.²

¹Faculdade de Engenharia Mecânica, Universidade Estadual de Campinas
Rua Mendeleyev, 200 - CEP 13083-860 - Campinas - SP
wariasr@fem.unicamp.br; wariasr@unal.edu.co

²Faculdade de Engenharia Mecânica, Universidade Estadual de Campinas
Rua Mendeleyev, 200 - CEP 13083-860 - Campinas - SP
wolf@fem.unicamp.br

Key words: Airfoil Noise, Trailing Edge Bluntness, DNS, Acoustic Analogy

Abstract. This work presents further results of the effects of trailing edge bluntness on airfoil tonal noise generation and propagation at low and moderate Reynolds numbers. Here, several simulations are conducted for a NACA 0012 airfoil at different free-stream Mach numbers. The effects of compressibility on sound generation and propagation are analyzed along with the effects of scattering by blunt trailing edges with different radii of curvature. Sound generation is computed by a hybrid method and an accurate two-dimensional direct calculation, and results are compared. The hybrid approach uses direct calculation for near-field source computations and the Ffowcs Williams-Hawkings equation as the acoustic analogy formulation. Numerical results show that the airfoil emits an intense “narrow-band” tone and that a thicker trailing edge emits higher noise levels than a thinner one since the magnitude of quadrupole sources are larger for the thicker configuration. At moderate Reynolds numbers, several tonal peaks are superimposed on a broadband spectrum for the thinner trailing edge investigated. For the blunter trailing edge, a tonal noise spectrum is shown.

1 Introduction

Due to the more stringent noise regulations and, since air traffic and wind power generation has increased, the understanding of trailing edge noise generation and propagation is an overriding concern for the design of low-noise aerodynamic shapes including wings and high-lift components, as well as wind turbine blades, propellers and fans. Brooks et al. [4] found that boundary-layer turbulence passing the trailing edge, vortex shedding due to laminar boundary layer instabilities and from blunt trailing edges are fundamental airfoil noise mechanisms, among other sources.
Paterson et al. [10] performed noise measurements from symmetric NACA airfoils with Reynolds number range between $10^5$ and $10^6$ at several angles of attack; their results showed the existence of discrete tones and multiple tones in a ladder-like structure pattern in terms of frequency and free-stream velocity. Arbey and Bataille [1] carried out experimental studies for three different NACA airfoils at 0 angle of attack, and showed that the tonal peak was a superposition of broadband contribution centered on a main frequency $f_s$ and a set of regularly spaced discrete frequency tones $f_n$. These authors also found the ladder-like structure from Paterson’s original work.

Nash et al. [9] performed experimental studies of airfoil tonal noise generation for a NACA 0012 profile with a Reynolds number of $1.45 \times 10^6$ and several angles of attack. They found a single dominant tonal frequency, instead of several peaks and no ladder-like structure of tonal frequency, in disagreement with the previous studies of Paterson et al. [10] and Arbey and Bataille [1]. Recently, Desquesnes et al. [5] performed direct numerical simulations (DNS) for a flow past a NACA 0012 airfoil for Reynolds numbers $1 \times 10^5$ and $2 \times 10^5$, and two angles of attack; their results showed the multiple tonal peaks. Tam and Ju [11] conducted DNS on a NACA 0012 airfoil for three different trailing edge thicknesses with a Reynolds number range of $2 \times 10^5$ to $5 \times 10^5$ at zero angle of attack. Their results showed only one tone for each simulation, in agreement with Nash et al. [9]. In addition, they suggested that an airfoil with a thicker trailing edge would have a lower tonal frequency for the same flow velocity. Hutcheson and Brooks [6] carried out experimental measurements of a cambered NACA varying the angle of attack, velocity and trailing edge bluntness concluding that, for increased air speed, the trailing edge increases noise levels and shifts noise to higher frequencies. Moreover, it was found that, for increased bluntness, the spectral peak increases in level and shifts to lower frequencies, in agreement with Tam and Ju [11].

The effects of trailing edge bluntness on generation and propagation of airfoil tonal noise at low Reynolds number were analyzed on a previous research [2]. This work continue the study of the effects of trailing edge bluntness on generation and propagation of airfoil tonal noise. 2-D DNS are performed for different rounded trailing edges of a NACA 0012 airfoil as shown in Fig. 1 and Tab. 1, at 3 deg. angle of incidence. Numerical results for Reynolds number based on the unitary airfoil chord $Re_c = 100000$ are showed for three different free-stream Mach numbers. The flow parameters of the selected cases are summarized in Table 2.

<table>
<thead>
<tr>
<th>Airfoil</th>
<th>Chord (c)</th>
<th>TE radius (r)</th>
<th>% r/c</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surface 1</td>
<td>0.98</td>
<td>0.0040</td>
<td>0.400</td>
</tr>
<tr>
<td>Surface 2</td>
<td>0.95</td>
<td>0.0060</td>
<td>0.600</td>
</tr>
<tr>
<td>Surface 3</td>
<td>0.90</td>
<td>0.0125</td>
<td>1.400</td>
</tr>
<tr>
<td>Surface 4</td>
<td>0.80</td>
<td>0.0250</td>
<td>3.125</td>
</tr>
</tbody>
</table>
Figure 1: Trailing edges profiles of the configurations analyzed.

Table 2: Characteristics of studied cases

<table>
<thead>
<tr>
<th>Reynolds $Re_c$</th>
<th>Mach number $M_\infty$</th>
<th>Mesh Points</th>
<th>$dz_0$</th>
<th>$dz_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5000</td>
<td>0.2 to 0.5</td>
<td>400×700</td>
<td>$5 \times 10^{-4}$</td>
<td>0.3</td>
</tr>
<tr>
<td>100000</td>
<td>0.1 to 0.3</td>
<td>400×900</td>
<td>$5 \times 10^{-4}$</td>
<td>0.3</td>
</tr>
</tbody>
</table>

$dz_0$ is the initial spacing and $dz_1$ is the end spacing of the mesh, in the wall normal direction. The O–mesh domain extends 45 chords from the airfoil in each direction.

2 Numerical Formulation

The general curvilinear form of the compressible Navier-Stokes equations is solved in conservation form. The numerical scheme for spatial discretization is a sixth-order accurate compact scheme [8] implemented on a staggered grid. Compact finite-difference schemes are non-dissipative and numerical instabilities arising from insufficient grid resolution, mesh non-uniformities and approximate boundary conditions have to be filtered to preserve stability of the numerical schemes. The high wavenumber compact filter presented by Lele [7] is applied to the computed solution at prescribed time intervals in order to control numerical instabilities. This filter is only applied in flow regions far away from solid boundaries. The time integration of the fluid equations is carried out by the fully implicit second-order scheme of Beam and Warming [3] in the near-wall region in order to overcome the time step restriction due to the usual near-wall fine-grid numerical stiffness. A third-order Runge-Kutta scheme is used for time advancement of the equations in flow regions far away from solid boundaries. No-slip adiabatic wall boundary conditions are applied along the solid surfaces and characteristic plus sponge boundary conditions are applied in the far field locations. The numerical tool has been previously validated for several simulations of compressible flows involving sound generation and propagation [13], [12], [2]. A hybrid approach which uses direct calculation for near-field source computations and the Ffowcs Williams-Hawkings (FW-H) equation as the acoustic analogy formulation is applied in the present work. Here, the 2-D FW-H equation is solved in the frequency domain and includes convective effects. The dipole and quadrupole sources are included in the formulation in order to perform an assessment of the individual sources on the far-field noise. More details about the numerical formulation can be found in reference [2].
3 Results

3.1 Flow configuration at $Re_c = 5000$

This section discusses results obtained by DNS and the FW–H acoustic analogy. In order to examine the effects of trailing edge bluntness on noise generation and propagation, hydrodynamic and acoustic properties are compared for airfoil surfaces No 1 (thinner trailing edge) and No 4 (blunter trailing edge). Figures 2 (a) – (d) present a detailed view of the mean flow streamlines at the trailing edge region, showing the development of recirculation bubbles for surfaces No 1 and No 4, for different free-stream Mach numbers and $Re_c = 5000$. One can see that, for the thinner trailing edge, the recirculation bubble is more pronounced on the airfoil suction side while for the thicker trailing edge, the recirculation region is more symmetric. From these figures, it is also possible to notice that flow separation occurs further upstream for higher Mach numbers and, therefore, recirculation bubbles become thicker and more elongated.

Figure 3 presents contours of pressure fluctuations for surfaces No 1 and No 4, at the vortex shedding frequency and $M_\infty = 0.2$ and 0.5, for $Re_c = 5000$. The contour levels are $10^{-5} < p' < 10^{-5}$ and $10^{-4} < p' < 10^{-4}$ for $M_\infty = 0.2$ and 0.5, respectively. From this figure, one can see that pressure fluctuations for surface No 4 (Figs.3 (c) and (d)) have higher amplitudes than those from surface No 1 (Figs.3 (a) and (b)) for wider angles of propagation, indicating that more acoustic energy is radiated from the airfoil with the blunter trailing edge. While the acoustic fields are considerably different, the near fields are identical with similar evanescent wave structures. The same results are found for the other free-stream Mach numbers studied.

In order to analyze the tonal noise structure emitted by the airfoils, Fig. 4 (a) presents
non-dimensional acoustic pressure spectra based on Helmholtz number $H_e=kc=2\pi f^*c^*$. Here, $f^*$ is the non-dimensional frequency and $c^*$ is the non-dimensional airfoil chord. Results are shown for surface No 1 for all Mach numbers investigated. One can observe a defined pattern of one dominant tonal peak besides the first harmonic for the higher Mach number cases. One can also notice that, for an increased Mach number, the dominant tonal peak increases in level and shifts to higher frequencies, in agreement with literature results. With the aim of quantifying the contribution of the individual noise sources, acoustic predictions are performed by the FW–H acoustic analogy for the flow configurations studied. Noise radiation from dipole and quadrupole sources are calculated and compared to direct noise calculations from DNS. Figure 5 shows the dipole and quadrupole source contributions to far field noise from surfaces No 1 and 4 at Mach numbers 0.2 and 0.5. Observer locations are at 10$c$ for Fig. 5 (a), and at 5$c$ for Fig. 5 (b). For the low Mach number case, dipole and quadrupole directivities resemble those of compact sources. When free stream Mach numbers are higher, dipoles and quadrupoles have directivities that depart from the classical shapes expected for compact sources since acoustic wavelengths are comparable to the airfoil chord. Moreover, effects of quadrupoles
become more pronounced and have a larger impact on the total far field noise as can be seen in Fig. 5 (b). From this figure, one can also notice that directivities obtained by acoustic analogy are found to be in excellent agreement with DNS results. Figures 6 (a) and (b) compare the DNS and FW–H directivities for all surfaces analyzed, for $M_\infty=0.2$ and 0.5, at the vortex shedding frequencies. It is observed that far field noise increases with the Mach number; furthermore, as trailing edge bluntness increases (for the same Mach number), the far field noise also increases.

Figure 7 shows the effect of trailing edge bluntness on the spatial distribution of quadrupole sources ($T_{11}, T_{12}$ and $T_{22}$). Here, solutions for the thinner (surface No 1) and the thicker (surface No 4) trailing edges are compared at the vortex shedding frequency for $M_\infty=0.2$. It is observed that, for the blunter trailing edge, quadrupole sources are more intense and have peak values closer to the solid surfaces. Furthermore, it is found in Fig. ?? that acoustic pressure amplitudes along the airfoil surface are five times larger for surface 4 than for surface 1, at $M_\infty=0.2$, and three times larger for surface 4 than that for surface 1 at $M_\infty=0.5$. These values of acoustic pressure represent the intensity of the dipole sources along the airfoil and one can conclude that the trailing edge acts as a powerful sound source since magnitudes of acoustic pressure are larger along these regions. As previously commented, if one considers the quadrupole sources as responsible by the incident sound field, and the airfoil surface as the sound scatterer, it can be concluded that a closer proximity of the former will increase the noise scattering by the latter (with a $1/\sqrt{r^5}$ algebraic sense for a two-dimensional acoustic wave). Here, $r$ is the distance between a point quadrupole source and a point on the solid surface. Moreover, larger magnitudes of the incident field will perform the same way.

### 3.2 Flow configuration at $Re_c = 100000$

This section discusses results obtained by DNS for the flow simulations at $Re_c = 100000$. In order to examine the effects of trailing edge bluntness on noise generation
Figure 5: Directivity plots for surfaces No 1 and 4 at the vortex shedding frequency. Observer locations at 10c for (a) and at 5c for (b).

Figure 6: Comparison of directivity plots at vortex shedding frequency for all trailing edge surfaces at 10c (a) and 5c (b).
and propagation, hydrodynamic and acoustic properties are compared for airfoil surfaces No 1 (thinner trailing edge) and No 4 (blunter trailing edge). Figure 8 presents pressure fluctuation spectra at near and far field for Mach numbers 0.1 and 0.3, for Surfaces 1 and 4. The blue lines are obtained at far field position and the red lines are computed on the hydrodynamic near field. It is clear from the figure that, for Mach = 0.1, the the tonal peaks are superimposed on the broadband spectrum, independently from the trailing edge thickness. For the higher Mach number, Mach = 0.3, one can see that the broadband spectrum is only present for the thinner trailing edge. In this case, the thicker trailing edge presented the typical tonal noise spectrum observed for lower Reynolds numbers.

In Fig. 9, one can see contours of magnitude of $T_{11}$ quadrupole sources for Surface 1. Figure 9 (a) shows the contours for $M_{\infty}=0.1$ and the low frequency tonal peak, outside of broadband hump in Fig. 8(a). In Fig. 9 (b), one can see the contours for $M_{\infty}=0.1$ and the maximum amplitude tonal peak frequency, inside the broadband hump in Fig. 8(a). One can see that quadrupole sources are more intense for the lower frequency shown in Fig. 8(a), even though the near field pressure fluctuation is lower. The near field pressure fluctuation is maximum for the main tonal peak inside the broadband hump. However, for this frequency, noise generation is less efficient. Similar trends are observed for Figs. 9 (c) and (d), obtained for $M_{\infty}=0.3$ and Surface 1.

Figure 10 (a) and (b) present contours of magnitude of $T_{11}$ quadrupole sources for Surface 4 at $M_{\infty}=0.1$ and 0.3, respectively. Both plots are obtained for the frequencies of maximum pressure fluctuation amplitudes in Figs. 8(c) and (d). It is possible to observe that, for Fig 10 (a), at $M_{\infty}=0.1$, the spatial distribution of quadrupole sources is intense behind the blunt trailing edge, indicating strong fluctuations of Reynolds stresses in that

Figure 7: Contours of magnitude of quadrupole sources, at the vortex shedding frequency, for $M_{\infty}=0.2$. 

(a) $T_{11}$ Surface No 1.  
(b) $T_{12}$ Surface No 1.  
(c) $T_{22}$ Surface No 1.  
(d) $T_{11}$ Surface No 4.  
(e) $T_{12}$ Surface No 4.  
(f) $T_{22}$ Surface No 4.
Figure 8: Spectra of pressure fluctuations at near and far field positions. Near field quantities shown in red and far field shown in blue.
(a) $M_\infty$=0.1 and low frequency tonal peak, outside of broadband hump in Fig. 8(a).

(b) $M_\infty$=0.1 and maximum amplitude tonal peak frequency, inside the broadband hump in Fig. 8(a).

(c) $M_\infty$=0.3 and low frequency tonal peak, outside of broadband hump in Fig. 8(b).

(d) $M_\infty$=0.3 and maximum amplitude tonal peak frequency, inside the broadband hump in Fig. 8(b).

**Figure 9:** Contours of magnitude of $T_{11}$ quadrupole sources for Surface 1.

(a) $M_\infty$=0.1 and maximum amplitude tonal peak frequency, inside the broadband hump in Fig. 8(c).

(b) $M_\infty$=0.3 and maximum amplitude tonal peak frequency in Fig. 8(d).

**Figure 10:** Contours of magnitude of $T_{11}$ quadrupole sources for Surface 4.
region. This is typical of blunt bodies. In Fig 10 (b), at \( M_\infty = 0.3 \), the spatial distribution of quadrupole sources extends throughout the wake region and, also, along the airfoil suction side, close to the trailing edge.

4 Discussion

The present work provides a systematic analysis of airfoil noise generation at low and moderate Reynolds numbers. The effects of trailing edge bluntness on noise generation and propagation over a NACA 0012 airfoil with different blunt trailing edge geometries are investigated for low to moderate Mach numbers. Results show that, for increased Mach number, the dominant tonal peak increases in amplitude and shifts to higher frequencies. For increased trailing edge bluntness, the dominant tonal peak increases in amplitude and shifts to lower frequencies. Furthermore, it is found that, a blunter trailing edge surface emits more noise than a thinner one due to an increase in the magnitudes of quadrupole sources near the trailing edge region. It is also found that the peak values of the quadrupole sources get closer to the airfoil surface for blunter trailing edges for the low Reynolds number investigated, which also increases noise scattering. For the moderate Reynolds number analysed, the typical summation of broadband and tonal contributions are observed, depending on the flow Mach number and trailing edge bluntness. It is found that airfoil noise generation for the main tonal frequency superimposed in the broadband hump is not the most efficient for the thinner trailing edge.

5 Acknowledgements

The authors acknowledge the financial support received from Fundação de Amparo à Pesquisa do Estado de São Paulo, FAPESP, under grants No. 2014/10166-6, No. 2013/03413-4 and No. 2013/07375-0, and the financial support received from Conselho Nacional de Desenvolvimento Científico e Tecnológico, CNPq, under Grant No. 470695/2013-7. The authors also thank to Centro Nacional de Processamento de Alto Desempenho em São Paulo, CENAPAD-SP, for providing the computer resources for the numerical simulations under project 551. The authors also acknowledge Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, CAPES, for providing a MSc. scholarship to the first author.

REFERENCES


ON THE DEVELOPMENT OF LARGE EDDY SIMULATION TOOLS FOR COMPRESSIBLE JET FLOW CONFIGURATIONS

Carlos JUNQUEIRA-JUNIOR\textsuperscript{1,2}, Sami YAMOUNI\textsuperscript{2}, João L.F. AZEVEDO\textsuperscript{3}

\textsuperscript{1,2}Instituto Tecnológico de Aeronáutica
DCTA/ITA, Computer Sciences Division
12228–900, São José dos Campos, SP, Brazil
\textsuperscript{1}junior.hmg@gmail.com \quad \textsuperscript{2}sami.yamouni@gmail.com

\textsuperscript{3}Instituto de Aeronáutica e Espaço
DCTA/IAE/ALA, Aerodynamics Division
12228–904, São José dos Campos, SP, Brazil
joaolhuiz.azevedo@gmail.com

Key words: LES, Compressible Flows, Jet Flows, Jet Noise Generation

Abstract. The current work addresses the numerical study of unsteady turbulent compressible flows for aeroacoustic applications. A novel large eddy simulation tool is developed in order to reproduce high fidelity results of compressible jet flows which are used for aeroacoustic studies using the Fowcs Williams and Hawkings approach. The numerical solver is an upgrade of a Reynolds-averaged Navier-Stokes code previously developed. Generally, large eddy simulations demand very dense grids. Hence, high performance computing is a requirement for such simulations. Therefore, the original framework is rewritten in Fortran 90 standards, and dynamic memory allocation and parallel computation have been added to the code. The communication between processors is performed by message passing interface protocols (MPI). The large eddy simulation formulation is written using the finite difference approach. Inviscid numerical fluxes are calculated using a second-order accurate centered scheme with the explicit addition of artificial dissipation. A five-step second-order accurate Runge-Kutta scheme is the chosen time marching method. The classical Smagorinsky model is the chosen turbulence closure for the present work. Numerical simulations of perfectly expanded jets with Mach number equal to 1.37 and Reynolds number of the jet equal to $5 \times 10^5$ are performed and presented here. Time averaged results are compared with numerical solutions and experimental data from the literature. Moreover, the speedup and the computational performance of the code are evaluated and discussed.
1 INTRODUCTION

The noise generated by the interaction of exhaust gases from rocket engines with the ambient air is a problematic topic since the 1950’s with the birth of the jet aircraft industry. The high amplitude acoustic waves generated by the supersonic jet have a critical impact on the environment, disturbing civilians comfort living near the airports for example. Moreover, they can be detrimental to the rocket itself, by damaging its structure. To control this phenomenon, it still remains necessary to better understand the driving physical mechanisms. In the literature, three main mechanisms of supersonic jet noise generation have been identified: the noise due to turbulent mixture processes, the noise due to shock waves and the well-defined tonal noise or screech noise. In the present work, since perfectly expanded jets are being considered, only the former mechanism is taken into account. It is commonly accepted that the mechanism of noise emission of this type can adequately be described by hydrodynamic instability theory, in which one can assume a slow evolution of the mixing layer [1]. Lighthill, through acoustic analogy, [2] was the first to identify the aerodynamic turbulent motions as the primary noise source. Later, Williams, Hawkings [3] and Lilley [4] extended the theory of Lighthill including mean flow convection and refraction effects. This theory has then be refined by experimental studies such as [5], [6], [7] or [8]. But it is only recently, with the advances of computer performances, that numerical jet noise prediction using with large-eddy simulation (LES) and direct numerical simulation (DNS) has succeeded [9, 10, 11]. At high Reynolds number, LES is the method of choice for the calculation of the flows of interest, since this formulation can capture accurately the most energetic scales associated with noise generation, being less expensive than DNS. However, this type of numerical method requires very refined mesh and the parallelization capability needs then to be implemented. Therefore, through laminar simulations, the objective of the present work is to prepare the computational tool which will yield accurate aerodynamic results to be used in future work to study the sound radiation due to a supersonic jet flow. In the latter, the acoustic analogy formulation of Ffowcs Williams and Hawkings (FWH) [3, 12] will be used. With this aim in mind, the authors are working on a new parallel LES tool at the Instituto de Aeronáutica e Espaço (IAE), in Brazil.

2 NUMERICAL METHOD: REYNOLDS AVERAGED NAVIER-STOKES SIMULATIONS

2.1 Governing equations

The Reynolds-averaged Navier-Stokes approach is based on time filtering procedure of the original Navier-Stokes equations. The high frequency fluctuations of the properties are filtered and the mean flow properties are calculated. Turbulence models, based on mean flow properties, are used in order to recover the contributions of the filtered properties on
the flow. The resulting time filtered flow properties can be written as

\[
\frac{\partial \rho}{\partial t} + \frac{\partial }{\partial x_j} (\rho \tilde{u}_i) = 0, \tag{1}
\]

\[
\frac{\partial }{\partial t} (\rho \tilde{u}_i) + \frac{\partial }{\partial x_j} (\rho \tilde{u}_i \tilde{u}_j) + \bar{p} \frac{\partial \tilde{u}_i}{\partial x_i} - \frac{\partial }{\partial x_j} \tau_{ij} = 0, \tag{2}
\]

\[
\frac{\partial e}{\partial t} + \frac{\partial }{\partial x_j} \left[ (e + \bar{p}) \tilde{u}_j - \tilde{u}_i \tau_{ij} + \overline{q_j} \right] = 0. \tag{3}
\]

where \((\cdot)\) and \((\ldots)\) are mean flow properties and Favre filtered properties, respectively. \(t\) and \(x_i\) are independent variables representing time and spatial coordinates of a Cartesian coordinate system, \(x\), respectively. The components of the velocity vector, \(\mathbf{u}\), are written as \(u_i\) for \(i = 1, 2, 3\). Density, pressure and total energy per unit of mass are denoted by \(\rho\), \(p\) and \(E\), respectively.

The effect of turbulence is included into the mean flow by a modification of the dynamic viscosity coefficient, \(\mu\), as \(\mu = \mu_l + \mu_t\). In this equation, \(\mu_l\) is the molecular dynamic viscosity coefficient, a property of the fluid that can be calculated by the Sutherland law, and \(\mu_t\) is a property of the flow and needs a turbulent closure to be calculated. Therefore, the components of the shear-stress tensor are given by

\[
\tau_{ij} = 2 (\mu_l + \mu_t) \left( \tilde{S}_{ij} - \frac{1}{3} \delta_{ij} \tilde{S}_{kk} \right), \tag{4}
\]

in which the averaged components of the rate-of-strain tensor, \(\tilde{S}_{ij}\), are written as

\[
\tilde{S}_{ij} = \frac{1}{2} \left( \frac{\partial \tilde{u}_i}{\partial x_j} + \frac{\partial \tilde{u}_j}{\partial x_i} \right). \tag{5}
\]

The thermal conductivity coefficient is also written in the same fashion, \(\kappa \rightarrow \kappa_l + \kappa_t\), where

\[
\kappa_t = \frac{\mu_tC_p}{Pr_t}, \tag{6}
\]

and \(Pr_t\) is the turbulent Prandtl number which, in the present work, is assumed as \(Pr_t = 0.9\). The time filtered heat flux, \(\overline{q_j}\), is given by

\[
\overline{q}_j = (\kappa_l + \kappa_t) \frac{\partial \tilde{T}}{\partial x_j}. \tag{7}
\]

In order to close the system of equations the density, the static pressure and the static temperature are related by the equation of state given by \(\bar{p} = \bar{p}R\tilde{T}\), where \(R\) is the gas constant, written as \(R = C_p - C_v\), and \(C_v\) is the specific heat at constant volume. The total energy per unity of mass is given by:

\[
\bar{p} \tilde{E} = \frac{\bar{p}}{\gamma - 1} + \frac{1}{2} \bar{p} \tilde{u}_i \tilde{u}_i, \tag{8}
\]

where \(\gamma\) is the ratio of specific heats, written as \(\gamma = C_p/C_v\).
2.2 Flow configuration and Mesh generation

In the present work, the case of interest is a perfectly expanded supersonic jet. The flow is characterized by a Mach number of 1.37, to be comparable to reference work such as [13], [14], [6], [9] and [7]. The chosen geometry is shown in Fig. 1 where a structured 2-D mesh of the supersonic jet is shown. The complete 3-D mesh is then created by rotating the 2D-mesh around the horizontal direction $x$. The authors chose not to include the nozzle geometry and the jet entrance is located at $x = 0$, between $|r|/D \leq 0.5$, where $D$ is the incoming jet diameter. In Fig. 1 two distinct regions are visible separated by a red dashed line: the developing region for $x/D < 30$ and the fully-developed region where the flow is self-similar. In the latter, the shear layer has a constant growth rate given by $S \approx 0.094$ [15]. This trend is typical of jet flows and the mesh refinement follows this evolution. The refinement is made using hyperbolic tangent functions, with a finest grid resolution near the jet entrance and along the slip line of the jet. As the code is fully compressible, the mesh is coarsened in the far field to diffuse the acoustic waves and avoid reflections on the farfield domain boundaries. Finally, as the solver is parallelized, the 2-D mesh is partitioned equally in the axial direction in order to keep a balanced computational cost between all the processors.

![Figure 1: 2-D view of the grid used for compressible jet flow simulations.](image)

3 HIGH PERFORMANCE COMPUTING

Parallel computation can largely decrease the time of CFD simulations and is necessary for large dimensional problem. In the current work, the parallel upgrade of a serial solver is performed ([16]). The parallel version of this solver ought to reproduce exactly the same results as the serial version. The computational performance of the parallel solver is measured and compared with the ideal case by using two main parameters: the speedup and the performance. The speedup $[17, 18, 19]$, $S_{p_N}$, is calculated as the ratio of the time spent to perform one thousand iterations of a flow simulation using $N$ processors, $T_N$, and
the time spent for the same configuration using a single processor, $T_s$, as presented in Eq. (9).

$$Sp_N = \frac{T_N}{T_s}. \quad (9)$$

The performance of the code using $N$ processors, $\eta_N$, is obtained by the ratio of measured speedup and the ideal speedup calculated by Amdahl’s law [20], and is given by

$$\eta_N = \frac{Sp_N}{N}. \quad (10)$$

In order to assess the code parallelization, the zero-pressure gradient flow over a one meter flat plate has been chosen. The Mach number is 0.3 and the Reynolds number over the flat plate is set to $6.641 \times 10^6$ for all the simulations [16]. Gustafson [18] states that the performance of a parallel task scales with the size of a problem. Therefore, different meshes were created in order to evaluate the performance of the solver for different number of points within partitions, $n_x = N_x/N$. Table 1 presents the mesh size and the number of points in the three directions of the domain. Figures 3 and 3 illustrate the trending evolution of speedup and performance as a function of the number of processors. One can observe that the code has the best computational performance for $n_x > 50$. The performance is $O(1)$ for all configurations with $n_x > 50$ presented in this section. For $50 > n_x > 40$ the communication costs become significant and begins to deteriorate the computational performance which decreases with smaller partitions size. For $n_x \approx 5$ the

\[\text{Figure 2: Trending of speedup on the left and performance on the right as a function of the number of processors.}\]
Table 1: Mesh information

<table>
<thead>
<tr>
<th>Mesh</th>
<th>Nx</th>
<th>Ny</th>
<th>Nz</th>
<th>Mesh size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>200</td>
<td>150</td>
<td>40</td>
<td>1.20E06</td>
</tr>
<tr>
<td>2</td>
<td>400</td>
<td>150</td>
<td>40</td>
<td>2.4E06</td>
</tr>
<tr>
<td>3</td>
<td>800</td>
<td>150</td>
<td>40</td>
<td>4.80E06</td>
</tr>
<tr>
<td>4</td>
<td>400</td>
<td>75</td>
<td>40</td>
<td>1.20E06</td>
</tr>
<tr>
<td>5</td>
<td>400</td>
<td>300</td>
<td>40</td>
<td>4.80E06</td>
</tr>
<tr>
<td>6</td>
<td>400</td>
<td>150</td>
<td>10</td>
<td>6.00E05</td>
</tr>
<tr>
<td>7</td>
<td>400</td>
<td>150</td>
<td>60</td>
<td>3.60E06</td>
</tr>
<tr>
<td>8</td>
<td>800</td>
<td>300</td>
<td>120</td>
<td>2.88E07</td>
</tr>
<tr>
<td>9</td>
<td>200</td>
<td>75</td>
<td>40</td>
<td>6.00E05</td>
</tr>
<tr>
<td>10</td>
<td>200</td>
<td>75</td>
<td>20</td>
<td>3.00E05</td>
</tr>
</tbody>
</table>

Communication is more expensive than computation. The speedup coefficient is higher for configurations with \( nx \approx 10 \) than configurations with \( nx \approx 5 \).

## 4 PERFECTLY EXPANDED SUPersonic JET FLOW

The present section is devoted to a preliminary study of a supersonic perfectly expanded jet flow. Results are compared with analytical, numerical and experimental data from the literature [13, 14, 6, 9, 7]. The flow is characterized by a perfectly expanded inlet jet with a Mach number of 1.37 at the jet exit. Only a viscous simulation without SGS filtering and turbulence modeling is included in the current extended abstract. LES results using the Smagorinsky SGS turbulence closure [21] will be discussed in the final paper. In the present work, the jet is considered as perfectly expanded and unheated. Therefore, the pressure ratio, \( PR = P_{jet}/P_{\infty} \), and the temperature ratio, \( TR = T_{jet}/T_{\infty} \), between the jet exit and the ambient freestream are equal to one, \( PR = 1 \) and \( TR = 1 \). The time step used in the simulation is constant and equal to \( 2.10^{-3} \) in dimensionless form. The Reynolds number of the jet is \( Re = 5 \cdot 10^5 \). The radial and longitudinal dimensions of the smallest cell of the computational domain are given by \( (\Delta r)_{min} = 0.03 \) and \( (\Delta x)_{min} = 0.064 \), respectively, again in dimensionless form. This cell is located in the shear layer of the jet, at the entrance of the computational domain. The number of points in the azimuthal direction is \( N_\theta = 180 \). The mesh domain is composed by 2.5 million points.

Serial and parallel simulations are performed in order to compare the convergence history and evaluate the transparency of the solver. Figure 3 illustrates the evolution of the continuity equation residue through one hundred iterations for serial and parallel simulations of the supersonic jet flow. Both simulations have visually presented the same convergence history. Therefore, one can accept that the solver modifications, in the context of parallelization of the code, are well implemented.

Generally, compressible jet flow studies performed in the literature present mean field
properties. Therefore, in the current work, the solution of the simulations are time-averaged. Mean fields are calculated through time integration using 48 field snapshots extracted at each 25000 iterations. The first time steps containing the transition phase of development of the jet are removed from the averaging process in order to obtain the cleanest time-averaged results possible.

Figure 4 presents a comparison between the streamwise velocity component profiles at different longitudinal locations of the current simulation, and the streamwise velocity component profile at \( x = 2.5 \) of the LES simulation of Mendez [13] on the left. First, one should note that the velocity profile at \( x = 0 \) has a square shape. This velocity profile is related to the perfect slipping velocity profile boundary condition fixed at the nozzle exit. The centerline values of the velocity profile at different positions are overestimated in the current simulation when compared with the results of Mendez. The difference is expected since the simulation performed in the present work is preliminary and neglects several important effects of the flow. However, one can observe a very good agreement between the current study and numerical and experimental results from Bodony [9] and Lau [22] for the centerline velocity as one moves downstream along the jet, as indicated in Fig. 4 on the right. The potential core length, corresponding to the jet area until \( U = 0.95 \cdot U_{\text{jet}} \), is about \( 19 \cdot r_0 \), which is in agreement with the literature [15]. The decrease of jet velocity as the jet is moving away from the “nozzle” is also illustrated in Fig. 6. In this figure, various slices are extracted at different positions along the longitudinal axis. The potential core area practically matches the red region of the jet. After this region, the flow enters in the self-similar regime of the mixing region, for \( x > 10D \). This self-similarity trend is observed in Fig. 5, where normalized velocity profiles at different streamwise locations are presented. The streamwise velocity component is normalized by its axial value \( U_c \). A reasonable collapse is obtained, in agreement with the work of Rona [23] at \( M = 2.0 \). The agreement improves when moving away from the potential core for \( x = 15D \) and \( x = 20D \). In this next region, the jet is supposed to display a linear spreading, consistent with the self-similarity.

A snapshot of the field in the permanent regime is taken and analyzed. Figures 4 and 4 display the instantaneous streamwise, radial and azimuthal velocity components of the supersonic jet, as well as its energy, density and pressure. One can observe that the jet dissipates before the end of the computational domain, at \( x \approx 150 \). This can be explained by the natural diffusion of the jet as well as the dissipation induced by the coarsening of the grid in the streamwise direction. Moreover, the flow does not display any discontinuity at the centerline and it seems to be unaffected by the domain external boundaries. Those observations indicate that the implemented boundary conditions are adequate for the present formulation without the use of turbulence closures. However, as previously stated, more sophisticated boundary conditions are necessary in order to achieve high fidelity LES results.

Finally, in Fig. 4, one can observe that the zone located downstream of the potential core exhibits high pressure level structures and such structures broadcast intense acoustic
Figure 3: Convergence histories obtained for serial and parallel simulations of the supersonic jet flow at $M = 1.37$ without turbulence modeling.

Figure 4: Comparison between streamwise velocity component profiles with LES results of Mendez [13] on the left and a comparison between time-averaged streamwise centerline velocity component with the work of Bodony [9] and Lau [22] on the right.

radiation. This observation is in agreement with Laufer [6]. Furthermore, the statement can also be verified in Fig. 8, where the jet velocity field is displayed in color superimposed upon the pressure field in gray scale. Even if this is not a time-averaged result, the source of pressure radiation is easily located after the jet potential core, corresponding to the high velocity region, again colored in red. The strong downstream directivity of these waves has already been observed in the literature [13].
Figure 5: Time-averaged velocity profiles from the jet flow at $M = 1.37$.

Figure 6: Mean streamwise velocity component for the present simulation at $M = 1.37$. The contours are colored from 0.1(black) to 1.4(red).

5 CONCLUDING REMARKS

The current study presents preliminary simulation results of the perfectly expanded supersonic jet flow. The parallelization of the code has been successfully performed. Dynamics of the supersonic jet has been well captured by the laminar simulation showing the same trends observed, in the reference work, regarding the Mach wave sources and their directions. At the moment, the turbulent simulations are converging while the large-eddy simulation tool is being implemented into the numerical tool. Future work will focus on the implementation of a fully turbulent inlet boundary layer, to reproduce the experimental unsteady dynamics of the jet, and on the implementation of the large-eddy simulations equations in the computational tool. The objective is to use the results to predict the acoustic radiation arising from the perfectly expanded supersonic jet flow.
Figure 7: Instantaneous properties of the perfectly expanded supersonic jet flow. Cartesian velocity components on the left and total energy, density and pressure on the right.

Figure 8: Instantaneous velocity magnitude and pressure fields for the present case at $M = 1.37$. The colored contours show the velocity magnitude from 0.1(blue) to 1.4(red), while the grey scale contours show $\rho/\rho_\infty$ from 0.7 to 0.72.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the partial support for this research provided by Conselho Nacional de Desenvolvimento Científico e Tecnológico, CNPq, under the Research Grants No. 309985/2013-7, No. 400844/2014-1 and No. 443839/2014-0. This work is also supported by Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, CAPES, through a Ph.D. Scholarship for the first author. The authors are also indebted to the partial financial support received from Fundação de Amparo à Pesquisa do Estado de São Paulo, FAPESP, under the Research Grants No. 2008/57866-1, No. 2013/07375-0 and No. 2013/21535-0.
REFERENCES


FLIGHT DYNAMICS PARAMETER ESTIMATION OF A ROTARY WING AIRCRAFT USING THE OUTPUT ERROR MINIMIZATION WITH NATURAL AND META HEURISTIC METHODS

Ivana Y. Sumida¹, Haroldo F. de Campos Velho¹, Eduardo F. P. Luz¹, Ronaldo V. Cruz², Luiz Carlos S. Góes²

¹Instituto Nacional de Pesquisas Espaciais (INPE)
Av. dos Astronautas, 1758 CEP: 12.227-010 São José dos Campos-SP, Brazil
E-mail: ivanayoshie@yahoo.com.br, haroldo@lac.inpe.br / URL: http://www.inpe.br

²Instituto Tecnológico de Aeronáutica (ITA)
Praça Mal. Eduardo Gomes, 50 - CEP: 12228-901 São José dos Campos-SP, Brazil
E-mail: goes@ita.br / URL: http://www.ita.br

Key words: Flight Simulation, Parameter Identification, Multiple Particle Collision Algorithm (MPCA)

Abstract. Flight simulators are employed by civil and military pilots, as well as engineers, in order to increase the security in training of crew, and to find out the behavior of the aircraft under different operational conditions. However, it is necessary to calibrate the simulator software to have good adherence to real flight. In this process, parameters of the mathematical model of the flight simulation need to be identified, such that the simulation is as close as possible to the real flight dynamic. With appropriated values of these parameters, the simulator will be ready for training or assessing the aircraft dynamics. This can be described as an inverse problem or parameter identification, formulated as an optimization problem. The simulator is designed to represent the dynamics of the helicopter AS355-F2, for testing two types of maneuvers were employed: a sinusoidal input and 3-2-1-1 pulse input. The aerodynamic derivatives estimation methodology is also known as quad-M scheme, since it involves four different processes: Measurement, Maneuver, Model, and Methods of error minimization. The tested helicopter was equipped with the Aydin Vector Data Acquisition System (AVDAS) PCU-816-I, ATD-800 digital recorder The system measures a total of thirty-five different parameters. The calibration of a dynamic flight simulator is achieved by two meta-heuristics: a Genetic Algorithm and a new approach named Multiple Particle Collision Algorithm (MPCA). Preliminary results show a good performance of the employed optimization methods.
INTRODUCTION

The Helicopter flight is quite expensive as compared with other similar fixed wing aircrafts. Helicopter Flight Simulators (HFS) can provide a suitable alternative to real flight experience to increase the flight security through training of the crew, prior evaluation of flight tasks and data acquisition procedures to validate and certificate onboard aircraft systems. However, the HFS must convey high degree of realism in order to be truly effective.

The area of parameter estimation and model identification have several applications in astronomy, aerospace, economics, biology, electrical, geological areas [1], [2], [3]. The strategy is to adjust the unknown model parameters in order to achieve the best fit between the predictions of the mathematical model and the experimentally observed system response. Tools and techniques of system identification have evolved to match the complexity of the models and the increasing need for correction and precision in the results. This methodology is more accurate than the corresponding values predicted by other methods such as analytical and numerical differentiation, [4], [5], [6].

The presence of noise such as state noise or measurement noise, [6], affect the identification methods, and this process becomes more difficult as the number of degrees of freedom (DOFs) and model parameters increase [7]. The identification of parameters methodology uses techniques such as the maximum likelihood method, equation error method, output error method, filter error method, and stochastic method [8], [9]. These methods require a mathematical model of the aircraft with a set of initial values for the parameters to begin the algorithm ([10], [11]). Concerning helicopter system identification techniques, very few articles have used stochastic method. One can cite [12], [13] in the longitudinal mode system identification of the Twin Squirrel helicopter and [14] in the identification of a small unmanned helicopter model. In this work, we used a well-known Quad-M methodology for parameters identification [15] as shown schematically in Figure 1.

![Figure 1: Adaptation Quad-M Method with MPCA.](image-url)
This methodology takes into account the main elements of rotorcraft system identification, including the rotorcraft excitation maneuvers, the aerodynamic data measurements, the mathematical model of the helicopter equations of motion, and the parameter estimation methods used to minimize the predicted output-error between the proposed aerodynamic model and the real data. With respect to the optimization of the output error, the method used was MPCA. Each one of these four different processes will be discussed in the following sections.

2 THE QUAD-M METHODOLOGY

The Quad-M or M⁴ methodology [15] takes into account the main elements of aircraft system identification, including the aircraft excitation maneuvers, the aerodynamic response measurements, the mathematical model of the aircraft equations of motion, and the parameter estimation methods used to minimize the output-error between the predicted model response and the real data observation. Each one of these functional elements is discussed below.

2.1 Maneuvers

The dynamic response of the helicopter in flight is excited by the application of different control inputs to the cyclic and collective flight commands including pulse signals, step, doublet, multistep, sinusoidal, and 3-2-1-1 pulse sequence, among others. Hence, a wide variety of manoeuvres can be specified to excite the specific modes of the aircraft.

The choice of a proper flight test maneuvers, by shaping the excitation signals, is very important to minimize the uncertainties in the parameter estimation procedures and to maximize the information in the flight test data content. The optimization of the excitation signal can be realized from a priori knowledge of the initial dynamic parameters of interest. However, since there are no priori studies available for AS355-F2 helicopter, the experimental manoeuvres were specified applying conventional flight test procedures and taking into account flight safety constraints. Since this work focuses the determination of the lateral-directional flight derivatives at forward and level flight, special sequences of sharp-edge pulses known as the 3-2-1-1 using the lateral cyclic and pedal inputs were used to excite the dutch-roll mode with 80 kts indicated airspeed at 5,000 ft of altitude pressure.

The identification procedure used the 3-2-1-1 sequence with both lateral cyclic and pedal inputs, while the validation procedure utilized two different sequences, one with lateral cyclic inputs and pedal fixed and the other with pedal input and fixed cyclic control, [12].

2.2 Measurements

The helicopter flight test data was recorded with the Aydin Vector Data Acquisition System (AVDAS) PCU-816-I and the ATD-800 digital recorder, this system measures a total of thirty-five different flight parameters. Some of the measured data channels include fuel
quantity in each tank, nose boom static and dynamic pressures, external stagnation temperature, aerodynamic angle of attack (α) and sideslip (β), roll, pitch, and yaw angular rates (p, q, r), load factors, longitudinal (θ) and lateral (φ) aircraft body attitudes, nose heading, and the flight command deflections, comprised of collective, longitudinal and lateral cyclic, and pedal command deflections (δ_c, δ_b, δ_a and δ_p).

The earth axis speeds (u, v, w) were obtained with the aid of a Z12 Differential Global Positioning System DGPS from Astech, whose antenna is fixed in the top of the helicopter vertical fin. The DGPS and AVDAS data synchronization was done inserting a simultaneous event in both systems. The DGPS data is represented with the same AVDAS data sampling rate by means of linear interpolation procedure.

The wind direction and intensity were obtained comparing the body axis speeds with the aerodynamic speed from the flight-test air data system, mounted on a nose boom, at trim conditions. Consequently, the body axis speeds (u, v, w) are easily calculated adding wind vector to the Earth axis speeds [13].

2.3 The Helicopter Model

The helicopter equations of motion are derived from Newton second law for the rigid body translational and rotational degrees of freedom, and are given by [16], [17], and [18] as:

\[ X = m(\dot{u} - rv + qw) + mg \sin \theta \]
\[ Y = m(\dot{v} - pw + ru) - mg \cos \theta \sin \phi \]
\[ Z = m(\dot{w} - qu + pv) - mg \cos \theta \cos \phi \]
\[ L = I_{xx} \dot{\phi} - I_{zz}(\dot{r} + pq) - (I_{yy} - I_{zz}) q r \]
\[ M = I_{yy} \dot{\phi} - I_{zz} (r^2 - p^2) - (I_{zz} - I_{xx}) r p \]
\[ N = I_{zz} \dot{\theta} - I_{xx} (\dot{\phi} - q r) - (I_{xx} - I_{yy}) p q \]

where X, Y and Z represents the external force components (longitudinal, lateral and vertical); L, M and N are respectively, the roll, pitch and yaw moments; and I(index) stands for the moments and products of inertia of the aircraft rotating body. The kinematic relation for the pitch rate and roll rate about Y and X-axis are written as:

\[ \frac{d\phi}{dt} = p + q \sin \phi \tan \theta + r \cos \phi \tan \theta \]  
\[ \frac{d\theta}{dt} = q \cos \phi - r \sin \phi \]  

The helicopter equations of motion are nonlinear, but a meaningful analysis can be employed by converting them into linear differential equations, by considering only small
Ivana Y. Sumida, Haroldo F. de Campos Velho, Eduardo F. P. Luz, Ronaldo V. Cruz, Luiz C. S. Góes

perturbations about a trimmed equilibrium point (represented by subscript $0$) in the rotorcraft flight envelope. In matrix notation, a linearized dynamical model is given by [16], and [18],

$$
\begin{bmatrix}
\Delta \dot{u} \\
\Delta \dot{q} \\
\Delta \dot{\phi} \\
\Delta \dot{\rho} \\
\Delta r
\end{bmatrix} = 
\begin{bmatrix}
X_{\mu} & X_{\nu} & X_{\mu} - u_0 & -g \cos \theta_0 & X_{\mu} & X_{\mu} - u_0 & -g \cos \phi_0 \cos \theta_0 \\
X_{\nu} & X_{\nu} & X_{\nu} + u_0 & -g \cos \phi_0 \sin \theta_0 & X_{\nu} & X_{\nu} + u_0 & -g \sin \phi_0 \cos \theta_0 \\
0 & 0 & 0 & -g \sin \phi_0 \sin \theta_0 & 0 & 0 & g \sin \phi_0 \sin \theta_0 \\
0 & 0 & 0 & \cos \phi_0 \sin \theta_0 & 0 & 0 & \cos \phi_0 \sin \theta_0 \\
N_\mu' & N_\nu' & \sin \phi_0 \tan \theta_0 & 0 & N_\mu' & N_\nu' & \sin \phi_0 \tan \theta_0
\end{bmatrix} \begin{bmatrix}
\Delta u \\
\Delta w \\
\Delta \theta \\
\Delta \phi \\
\Delta \rho \\
\Delta r
\end{bmatrix} + 
\begin{bmatrix}
X_{\mu B} & X_{\nu B} & X_{\mu A} & X_{\nu A} & X_{\mu I} & X_{\nu I} & X_{\mu P} & X_{\nu P} & X_{\mu I} & X_{\nu I}
\end{bmatrix} 
\begin{bmatrix}
\delta \mu \\
\delta \nu \\
\delta \theta \\
\delta \phi \\
\delta \rho \\
\delta r
\end{bmatrix} + \dot{x}_{bias}
$$

therefore, Equation 9 may also be written as:

$$
d\begin{bmatrix} X_l \\ X_d \end{bmatrix} =
\begin{bmatrix} A_l & C_l \\ C_2 & A_d \end{bmatrix} \begin{bmatrix} X_l \\ X_d \end{bmatrix} + 
\begin{bmatrix} B_l & D_l \\ B_2 & D_d \end{bmatrix} \begin{bmatrix} \Delta \delta_l(t - \tau) \\ \Delta \delta_d(t - \tau) \end{bmatrix} + \dot{x}_{bias}
$$

here $X_l$ and $X_d$ represent the state space vector for the longitudinal and lateral movements, respectively.

In this work, it is considered the parameter estimation of the lateral dynamic motion, whose simplifies model is expressed by:

$$
\frac{dX_d}{dt} = A_dX_d + B_d\Delta \delta_d(t - \tau) + \dot{x}_{bias}
$$

$$
X_d = \begin{bmatrix} \Delta u \\ \Delta \rho \\ \Delta \phi \end{bmatrix}^T
$$

$$
\Delta \delta_d = \begin{bmatrix} \Delta \delta_u \\ \Delta \delta_\rho \\ \Delta \delta_\phi \end{bmatrix}^T
$$

The parameter values of interest for system identification are the elements of matrix $A_d$ (stability derivatives), matrix $B_d$ (control derivatives), and the delays associated with the aircraft response ($\tau$). Furthermore, the parameters include the estimation of an unknown bias vector, $x_{bias}$. This vector is introduced in the mathematical model to represent measurement errors and noise produced by transducers and signal condition instrumentation [4].
The model parameter vector ($\Omega$) is estimated by a minimization process of the cost function, $J(\Omega)$, related to the output error between the measured and predicted system response,

$$J(\Omega) = \sum_{i=1}^{n} \| X_i^{obs} - X_i^{mod(\Omega)} \|_2^2$$

(14)

$$\Omega = \left( \frac{Y}{m}, \frac{Y_0}{m}, \frac{Y_p}{m}, L'_v, L'_r, N'_v, N'_r, \frac{Y_0}{m}, \frac{Y_p}{m}, L'_0, L'_p, N'_0, N'_p, \right.$$  
$$\left. \Delta v_{bias}, \Delta p_{bias}, \Delta \phi_{bias}, \Delta v_{bias}, \Delta v_{ref}, \Delta p_{ref}, \Delta \phi_{ref}, \Delta v_{ref}, \alpha, \tau \right)$$

(15)

where $n$ is the number of observed measurements.

### 2.4 The MPCA Optimization Method

The cost function to be minimized is the output error between the model prediction response and the actual measured response. This objective function depends on the parameters of the proposed dynamic model, such as the helicopter aerodynamic stability and control derivatives, sensor bias, and sensitivities. Therefore, the determination of a parameter vector $\Omega$ that minimizes the cost function given by Equation 14 can be seen as an optimization problem and will be solved by a new meta-heuristics, named the Multiple Particle Collision Algorithm (MPCA).

The MPCA optimization algorithm was inspired on typical physical phenomena related to neutron particle transport inside a nuclear reactor core, where during the neutron travel multiple particles absorption and scattering are observed. The results obtained with MPCA in this study are compared to the ones obtained by Cruz ([13]) where a Genetic Algorithm (GA) was used to find the helicopter aerodynamic and control derivatives.

The MPCA is a meta-heuristic optimization method based on the canonical PCA [19]. This version uses multiple particles in a collaborative way, organizing a population of candidate solutions. The PCA was inspired by the traveling process (with absorption and scattering) of a particle (neutron) in the core of a nuclear reactor. The use of the PCA was effective for several test functions and real optimization applications [20].

The PCA starts with a selection of an initial solution (Old-Config), and is modified by a stochastic perturbation (Perturbation{.}), leading to the construction of a new solution (New-Config). The new solution is compared (function Fitness{.}), and the new solution can or cannot be accepted. If the new solution is not accepted, the scheme of scattering (Scaterring{.}) is used. The exploration around closer positions is guaranteed by using the functions Perturbation{.} and Small-Perturbation{.}. If the new solution is better than the previous one, this new solution is absorbed. If a worse solution is found, the particle can be sent to a different location of the search space, such that it enables the algorithm to escape from a local minimum [21].
The implementation of the MPCA algorithm is similar to PCA, but it uses a set with \( N \) particles, where a mechanism to share the particle information is necessary. A blackboard strategy is adopted, where the best-fitness information is shared among all particles in the process. This process was implemented in Message Passing Interface (MPI), looking for application into a distributed memory machine [21]. The pseudo-code for the MPCA is presented by Table 1.

Table 1: MPCA: pseudo-code for the optimization algorithm.

<table>
<thead>
<tr>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generate an initial solution: Old-Config</td>
</tr>
<tr>
<td>Best-Fitness = Fitness{Old-Config}</td>
</tr>
<tr>
<td>Update Blackboard</td>
</tr>
<tr>
<td><strong>For n = 0 to # of particles</strong></td>
</tr>
<tr>
<td>For n = 0 to # iterations</td>
</tr>
<tr>
<td>Update Blackboard</td>
</tr>
<tr>
<td>Perturbation{}</td>
</tr>
<tr>
<td>If Fitness{New-Config} &gt; Fitness{Old-Config}</td>
</tr>
<tr>
<td>If Fitness{New-Config} &gt; Best-Fitness</td>
</tr>
<tr>
<td>Best-Fitness = Fitness{New-Config}</td>
</tr>
<tr>
<td>End If</td>
</tr>
<tr>
<td>Old-Config = New-Config</td>
</tr>
<tr>
<td>Exploration{}</td>
</tr>
<tr>
<td>Else</td>
</tr>
<tr>
<td>Scattering{}</td>
</tr>
<tr>
<td>End If</td>
</tr>
<tr>
<td>End For</td>
</tr>
</tbody>
</table>

3. RESULTS

The computational results obtained with MPCA and GA are show in Figures 2, 3 and 4. The GA and the MPCA have been implemented in the Matlab/Simulink environment. Computer tests were conducted under Linux operating system, in an Intel Core i5 2.27 GHz. The sinusoidal maneuver is represented by \( \delta \) and the results presented take into consideration the average of 4 experiments with seeds generate with different random numbers and experimental data generating artificially. The parameters used are: 2 particles; 10 iterations (exploration). The stopping criterion used was the total number of iterations (30) and the initial estimative were the derivatives of stability and control of BO105 improved by the GA
Figure 2 shows the lateral velocity as function of time. The red curve corresponds to the real data obtained during the test, the dotted blue is the result of the identification produced by the GA, and the results achieved by the MPCA are represented by the dotted curve in magenta. The results show that a small discrepancy between the measured data and the data obtained by both algorithms.

![Figure 2: Linear velocity variation along the Y-axis.](image)

A similar behavior is observed for the roll rate and the bank angle attitude of the aircraft as shown by Figures 3 and 4.

![Figure 3: Angular velocity (roll rate) variation along the X-axis.](image)
3 CONCLUSIONS

In this work, we compared two stochastic algorithms, GA and MPCA, for helicopter parameter identification. The techniques were applied only in the estimation of the aerodynamic parameters of the lateral motion. The problem is formulated as an optimization process. Different heuristic search algorithms (GA and MPCA) were employed to address the solution of the optimization problem. The results indicate that GA and MPCA present a good agreement, but the results are better with the MPCA implementation. Further work is suggested to apply MPCA in lateral-directional dynamic mode and in a more complex model which includes both longitudinal and lateral-directional dynamic modes of the helicopter.

ACKNOWLEDGEMENTS

The authors would like to thank CNPq (INCT) and CAPES for the financial support.

REFERENCES


STeady and unsteady analysis of aerodynamics wing sections at ultra-low Reynolds numbers (Re < 10000)

Dino P. Antonelli¹, Carlos G. Sacco² and Jose P. Tamagno³

¹Universidad Nacional de Cordoba y CONICET
Av. Velez Sarfield 1611, PC 5000 Cordoba, Argentina
dinoantonelli@hotmail.com

²Instituto Universitario Aeronautico
Av. Fuerza Aerea S/n, PC 5000 Cordoba, Argentina
csacco@iua.edu.ar

³Universidad Nacional de Cordoba
Av. Velez Sarfield 1611, PC 5000 Cordoba, Argentina
jtamagno@efn.uncor.edu

Key words: aerodynamic wing sections, ultra-low Reynolds, CFD, steady and unsteady flows.

Abstract. The purpose of this study is to describe phenomena that manifest themselves in flows where Reynolds numbers are ultra-low (Re < 10000). To accomplish this study, mathematical techniques capable of solving the Navier-Stokes equations for laminar-incompressible flows are used. It is noted that a solver based on the Finite Element Method provides an appropriate resolution procedure, however, it must also be noted that because of the incompressible assumption the character of the continuity equation goes from hyperbolic to elliptic. Because of this, a Fractional Step method which evolves toward a semi-implicit temporal integrator is used, and to handle the convective and pressure terms the so called Orthogonal Sub-grid Scale (OSS) algorithm is applied. In addition, the motion of the finite elements computational mesh through solving the Poisson equation and optimizing each element metric, is implemented. Basic useful results describing the behavior of several 2D geometries at steady ultra-low Reynolds flows, are presented. Different geometric parameters like thickness ratio, mean lines camber, shape of leading edge, etc. are changed and its effects evaluated. Flow detachment features and their impact on main aerodynamic properties are assessed. Wing sections performing typical unsteady flights like heaving, pitching, flapping and hovering are also analyzed, and its aerodynamic properties in terms of Strouhal numbers, reduced frequencies and Reynolds numbers determined.
1 INTRODUCTION

The importance of ultra-low Reynolds flows lies in technology applications like MAVs (Micro Air Vehicles). Numerous cases require a deep understanding of the present phenomena in both steady and unsteady flights, to obtain maximum propulsive and handling efficiencies. Basically due to Reynolds number effects, aerodynamic characteristics such as lift, drag and thrust of a flight vehicle change considerably between MAVs and conventional manned air vehicles. In fact, in the nature, birds or insects flap their wings interacting with the surrounding air to generate lift to stay aloft or producing thrust to fly forward. The main powered flights are: flapping (flight with free stream) and hovering (flight without free stream).

Much research in this broad area have been made. The most significant that can be named are: Kunz [8] in his thesis studied the behavior of different geometries in steady flows at ultra-low Reynolds number; Guerrero [7] carried out unsteady aerodynamic studies at ultra-low Reynolds in 2D and 3D configurations built using the NACA 0012 wing section; Pedro et al [10] with the purpose of studying the propulsive efficiency of flapping hydro-foil NACA 0012 at $Re = 1100$, (flow density $\rho = 1 kg/m^3$ and dynamic viscosity $\mu = 0.01 kg/(m.s)$ ), also carried out numerical simulations. A Finite Volume Technique with an additional equation for the pressure, an explicit temporal scheme and structured grid, were used.

Ranges of non-dimensional numbers found relevant to unsteady flights of biological "flappers", are also considered valid for MAVs. A characteristic one for flapping motions is the Strouhal number $St = f_h h_a/U$, where $f_h$ is the frequency and $h_a$ the amplitude. Therefore, the Strouhal number expresses de ratio between the flapping wing velocity and the reference velocity $U$. The reduced frequency given by $k = \pi f_c / U$ is another parameter that can be interpreted as a measure of unsteadiness comparing the wave length of the flow disturbance to the chord $c$.

2D unsteady flow sinusoidal kinematics are characterized by the equations:

$$h(t) = h_a \sin(2\pi f_h t + \phi_h) \quad (1)$$

$$\alpha(t) = \alpha_a \sin(2\pi f_a t + \phi_\alpha) \quad (2)$$

where $\phi_h$ and $\phi_\alpha$ are the phases angles.

2 NUMERICAL SIMULATION

2.1 Governing equations

The two-dimensional time-dependent Navier-Stokes equations are solved using the finite element method, assuming incompressible-laminar flow which is justified since the Mach number of MAV flight is $M << 0.3$ and the Reynolds number $Re < 10000$. Conservation of mass and momentum are described by:
\[ \nabla \cdot \mathbf{u} = 0 \quad \text{in} \quad \Omega \times (t_0, t_f) \] (3)

\[ \frac{\partial (\mathbf{u})}{\partial t} + (\mathbf{u} \cdot \nabla) \mathbf{u} = -\frac{1}{\rho} \nabla p + \nu \nabla^2 \mathbf{u} + \mathbf{f}_e \quad \text{in} \quad \Omega \times (t_0, t_f) \] (4)

where \( \Omega \) represents the analysis domain with boundaries \( \Gamma_u \cup \Gamma_\sigma \), whereas \( (t_0, t_f) \) is time interval of analysis. The \( \mathbf{u} \) is the two-dimensional flow velocity vector, \( \rho \) the constant density, \( \nu \) the kinematic viscosity and \( p \) the pressure.

To represent the unsteady flow, eqs. 3 and 4 are solved in a fixed inertial reference frame incorporating a moving mesh following the Arbitrary Lagrangian Eulerian (ALE) formulation [6]. This method combines the advantages of both the Lagrangian and Eulerian approaches. In the Lagrangian approach the computational mesh is moved such that the nodes follow material particles during motion. The computational mesh is fixed and the fluid moves with respect to the mesh. The ALE method incorporates a moving mesh using the Lagrangian method, where the mesh follows the motion of the geometry boundary, whereas the equations are solved using the Eulerian approach.

In order to obtain the ALE equations the velocity \( \mathbf{u} \) in the convective term of the momentum equation needs to account for the mesh motion. Therefore the velocity of the mesh \( \mathbf{u}_m \) is subtracted from the flow velocity in the convective term. Then the Navier-Stokes equations in ALE formulation are obtained by:

\[ \nabla \cdot \mathbf{u} = 0 \] (5)

\[ \frac{\partial (\mathbf{u})}{\partial t} + (\mathbf{c} \cdot \nabla) \mathbf{u} + \frac{1}{\rho} \nabla p - \nu \nabla^2 \mathbf{u} - \mathbf{f}_e = 0 \] (6)

where \( \mathbf{c} = \mathbf{u} - \mathbf{u}_m \) is the convective velocity that represent the difference between fluid velocity and mesh velocity.

In the present work, the algorithm of mesh movement is based in operations of optimization of smoothing, developed to the @ANSYS software package [3].

### 2.2 Fractional Step algorithm

The equations previously presented can’t be solved by a numerical standard form because incompressibility gives raise to a flow field restriction. There are several algorithms to deal with this difficulty and the Fractional Step method is one of them. The method meets the LBB condition through the use of same order of approximation for velocity and pressure.

To apply the Fractional Step algorithm the momentum equation is divided in two parts:

\[ \hat{\mathbf{u}}^{n+1} = \mathbf{u}^n + \delta t \left[ \mathbf{u}^{n+\theta} \cdot \nabla \mathbf{u}^{n+\theta} + \gamma \frac{1}{\rho} \nabla p^n - \nu \nabla^2 \mathbf{u}^{n+\theta} + \mathbf{f}^{n+\theta} \right] \] (7)
\[ u^{n+1} = \hat{u}^{n+1} - \frac{\delta t}{\rho} (\nabla p^{n+1} - \gamma \nabla p^n) \] (8)

and in the last two equations a new variable \( \hat{u} \) known as fractionary momentum, is introduced. If in eq. 8 the divergence is taken and the continuity equation is applied, results:

\[ \nabla^2 (\nabla p^{n+1} - \gamma \nabla p^n) = \frac{\rho}{\delta t} \nabla \hat{u}^{n+1} \] (9)

Through this equation the pressure is calculated. In addition, \( \gamma \) is a numerical parameter such that its values of interest are 0 and 1. The \( \theta \) parameter determine the kind of temporal approximation.

### 2.3 Discret form of equations

The Finite Element Method is used to discretize the govern equations and provides an appropriate resolution procedure [9]. The resultant scheme is of first order (\( \gamma = 0 \)) and the temporal discretization (\( \theta = 0 \)) results in Euler forward. The test functions \( (\psi_h, \phi_h) \in \Psi_h \times \Phi_h \) are used such as 1:

\[
\frac{1}{\delta t} (\hat{u}_h^{n+1}, \psi_h) = \frac{1}{\delta t} (u_h^n, \nabla \psi_h) - (u_h^n, \nabla u_h^n, \psi_h) - \nu (\nabla u_h^n, \nabla \psi_h) - (f^n_e, \psi_h) \] (10)

\[
(\nabla p_h^{n+1}, \nabla \phi_h) = \frac{\rho}{\delta t} [ (\hat{u}_h^{n+1} - u_h^n, \nabla \phi_h) - (\nabla u_h^n, \phi_h) ] \] (11)

\[
(u_h^{n+1}, \psi_h) = (\hat{u}_h^n, \psi_h) - \frac{\delta t}{\rho} (\nabla p_h^{n+1}, \phi_h) \] (12)

The last equations system is semi-implicit because eqs. 10 and 12 are explicit (lumped mass matrix) and eq. 11 for the pressure computation is implicit.

### 2.4 Stabilized scheme

The discretization of convective terms yields numerical instabilities, therefore stabilization methods must be used. In this work the Orthogonal Subgrid Scale (OSS) algorithm is applied [4],[5],[11]. The expresion for the convective stabilization term is:

\[ STB_u = \tau_1 (u_h^n, \nabla u_h^n - \pi_h^n, u_h^n \nabla \psi_h) \] (13)

where \( \pi_h^n \) is the convective term proyecion and it is defined in eq. 18. This equation add to momentum eq. 10 and it is evaluated in \( t^n \), therefore it remains explicit.

The term stabilization of pressure to be added to the eq. 11 is:

1The notation used in the equations mean: \((a, b) = \int a \cdot b \, d\Omega\)
where \( \xi _{h}^{n} \) is the gradient pressure term projection and it is defined in eq. 19. In addition, it is evaluated in \( t^{n} \), therefore it remains explicit.

The complete stabilized scheme is obtained:

\[
\frac{1}{\delta t} (\hat{u}_h^{n+1}, \psi_h) = \frac{1}{\delta t} (u_h^n, \psi_h) - (u_h^n \cdot \nabla u_h^n, \psi_h) - \nu (\nabla u_h^n, \nabla \psi_h) - (f^n_h, \psi_h) - (\tau_1 (u_h^n \cdot \nabla u_h^n - \pi_h^n), \nabla u_h^n \cdot \nabla \psi_h)
\]

(15)

\[
(\nabla p_h^{n+1}, \nabla \phi_h) = \frac{\rho}{\delta t + \tau_2} \left[ (\hat{u}_h^{n+1} - u_h^n, \nabla \phi_h) - (\nabla u_h^n, \phi_h) \right] + \frac{\tau_2}{\delta t + \tau_2} (\nabla \xi_h^n, \nabla \phi_h)
\]

(16)

\[
(u_h^{n+1}, \psi_h) = (\hat{u}_h^n, \psi_h) - \frac{\delta t}{\rho} (\nabla p_h^{n+1}, \phi_h)
\]

(17)

\[
(\pi_h^n, \tilde{\psi}_h) = (u_h^n \cdot \nabla u_h^n, \tilde{\psi}_h)
\]

(18)

\[
(\xi_h^n, \tilde{\psi}_h) = (\nabla p_h^n, \tilde{\psi}_h)
\]

(19)

where \( \tilde{\psi}_h \in \tilde{\Psi}_h \). The system of equations of eqs. 15, 17, 18, 19 are solve in explicit form with lumped mass matrix and the system resultant of eq. 16 is solve in explicit form through of conjugate gradients with diagonal pre-conditioner.

It is noted that the formulation of the scheme isn’t in the ALE framework. To account the mesh velocity is necessary introduce in convective and stabilizations terms, the convective velocity \( c \).

Finally the boundary conditons in viscous tensor and velocity are:

- Imposed velocity: \( u = u_c \)
- No slip: \( u = 0 \)
- No traction: \( n.\bar{\sigma}.n = 0 \)

### 3 Verification of Numerical Code

For the cases studies included in Table 1 the following parameters are considered: pitching and heaving frequencies \( f_\alpha = f_h = 0.225[H \bar{z}] \), reducy frequency \( k = 0.7096 \), maximum heaving amplitude \( h_a = 1 \), phase angle \( \varphi = 90^\circ \), Strouhal number \( St = 0.45 \) and the variable parameter is the pitching amplitude.

From Table 1 it can be concluded that the results obtained in this work compare well with those given by [10] and [7], up to \( \alpha_a = 15^\circ \), but no so much for greater angles \( \alpha_a = 20^\circ \) and \( \alpha_a = 25^\circ \). In this cases computed values from this work tend to overpredict results given by the other autors. Specific investigations about the reason for differences were not made.
Table 1: Average thrust $\bar{ct}$ and maximum lift $C_{L_{\text{max}}}$ coefficients. Comparison for flapping NACA 0012 cases between present work, Refs. [10] and [7].

<table>
<thead>
<tr>
<th>$\alpha_0$</th>
<th>Pedro et al.[10] $ct$</th>
<th>Guerrero[7] $C_{L_{\text{max}}}$</th>
<th>Present Work $ct$</th>
<th>Present Work $C_{L_{\text{max}}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$5^\circ$</td>
<td>0.4324</td>
<td>8.3333</td>
<td>0.4245</td>
<td>8.0828</td>
</tr>
<tr>
<td>$10^\circ$</td>
<td>0.6511</td>
<td>7.4834</td>
<td>0.6576</td>
<td>7.1699</td>
</tr>
<tr>
<td>$15^\circ$</td>
<td>0.8226</td>
<td>6.6307</td>
<td>0.8360</td>
<td>6.5435</td>
</tr>
<tr>
<td>$20^\circ$</td>
<td>0.9337</td>
<td>5.8176</td>
<td>0.9389</td>
<td>6.1133</td>
</tr>
<tr>
<td>$25^\circ$</td>
<td>1.0046</td>
<td>5.0558</td>
<td>0.9601</td>
<td>5.6080</td>
</tr>
</tbody>
</table>

4 RESULTS AND DISCUSSION

4.1 Steady analysis of aerodynamic airfoils

Figure 1: Lift and drag coefficients for thickness ratios and camber ratios over standard NACA 4 digit airfoils. (a) and (b) thickness ratio effects. (c) and (d) camber ratio effects. (The inviscid curve is from [8]).

To analyze how the thickness ratio behaves in ultra-low Reynolds flow, comparisons between four digits symetric NACA airfoils (0002, 0006, 0008) at two Reynolds numbers ($Re = 2000$ and $Re = 6000$) are carried out. Results are presented in Fig. 1.a and 1.b. Consider first the 0006 and 0008 thickness airfoils. In the quasi-linear ideal range of angles
of attack ($0^\circ - 6^\circ$) for a symmetric airfoil (Fig.1.a), a very important reduction in lift can be observed. If the Reynolds number is decreased the lift slope tends to improve, but such improvement is comparatively less than the decrease due to thickness. However, in NACA 0002 airfoil the lift slope becomes greater when the Reynolds is higher ($R_e = 6000$). It appear that his behavior is related to a lower leading edge suction peak and subsequent delayed stall. Note from Fig. 1.a that as the thickness ratio get smaller, the closer the viscous results get to the ideal values. Fig. 1.b shows the strong increase of drag coefficients as the Reynolds number decreases.

To analyze the camber behavior a comparison between NACA 2302, 4302 and 6302 is performed. Lift results of numerical simulations plotted in terms of $\alpha - \alpha_0$, being $\alpha_0$ the zero lift angle, are shown in Fig. 1.c. The 6302 airfoil lift slope, is 30% greater than 2302 airfoil. If $C_L = 0$, note the large increment of $C_D$ as the camber increases Fig. 1.d, it can be attributed to leading edge early detachments.

4.2 Unsteady analysis of aerodynamic airfoils

Figure 2: Average thrust coefficient $\bar{c}t$ and propulsive efficiency $\eta$ for NACA airfoils (0004, 0006 and 0012) in heaving motion. Plots (a) and (b) apply to $f = 1[Hz]$. Plots (c) and (d) apply to $f = 2[Hz]$.

The fundamental parameter of unsteady analysis is the Strouhal number, defined as $St = 2fh_a/U$. Taylor et al. [12] and Triantafyllou et al. [13] performed a study of wing frequencies and amplitudes, and cruise speeds across a range of birds, insects, fishes and cetaceans, to determine Strouhal numbers in “cruising” flight. They found 75% of the 42
species considered fall within a narrow range of $0.19 < St < 0.41$ \cite{7}. Therefore, a similar range of Strouhal numbers has in this work been selected.

### 4.2.1 Heaving

The kinematic of heaving motion is given by eq. 1. The analysis is applied to NACA’s four digits (0004, 0006 and 0012) symmetrical airfoils. The kinematics parameters are: two values of heaving frequencies $f_h$ and a variable Strouhal number throughout the heaving amplitude $h_a$. The average thrust coefficient $\overline{ct}$ and propulsive efficiency $\eta = \overline{ct}/\overline{cp}$ (where $\overline{cp}$ is the power coefficient input), are in terms of the Strouhal number presented for $f_h = 1 [Hz]$ and $f_h = 2 [Hz]$ in Fig. 2.a and 2.b, and in Fig. 2c and 2d respectively. Note the numerical results obtained by Guerrero \cite{7} in Fig. 2.a and 2.b.

![Figure 3: Comparison at different times of velocity contours between NACA 0004 and NACA 0012 airfoils ($f_h = 1 [Hz]$ and $St = 0.3$). Times (a) and (e) $t = 0.45 [s]$, (b) and (f) $t = 0.86 [s]$, (c) and (g) $t = 1.29 [s]$, (d) and (h) $t = 1.64 [s]$.](image)

The flow motion topology expressed by Figure 3, helps to understand the results of the simulations. Therefore, comparisons at different times of velocity contours between NACA 0004 (Fig. 3 a,b,c,d) and NACA 0012 (Fig. 3 e,f,g,h) are shown. The formation of leading edge vortices (LEV) and its convection into the wake, can be observed \cite{1}.

### 4.2.2 Flapping

The name flapping is applicable to a combined motion of heaving and pitching, consequently the kinematics relations given by equations 1 and 2 are simultaneously applied.

In Fig. 4a and 4b are shown average thrust coefficients $\overline{ct}$ and propulsive efficiencies $\eta$ applicable to NACA symetric airfoils 0004 and 0012, as function of the pitching angle.
\[5^\circ < \alpha_a < 25^\circ\] with \(f_h = f_\alpha = 0.3, h_a = 0.5[m], \varphi = \pi/2\). In Fig. 4c and 4d are shown average thrust coefficients \(\bar{c}_t\) and propulsive efficiencies \(\eta\) applicable to NACA symmetric airfoils 0004 and 0012, as function of the heaving amplitude \(0.025 < h_a < 0.5\) with \(f_h = f_\alpha = 1, \alpha_a = 15^\circ\) and \(\varphi = \pi/2\).

**Figure 4**: Average thrust coefficient \(\bar{c}_t\) and propulsive efficiency \(\eta\) in heaving motion for thickness ratio variation in NACA 0004, 0012 at \(Re = 1100\). (a) and (b) pitching amplitude variable \(\alpha_a\). (c) and (d) heaving amplitude variable \(h_a\).

### 4.2.3 Hovering

The kinematics relations in hovering flight are given by equations 1 and 2 simultaneously applied. The kinematic parameters utilized in the simulation of hovering are: \(h_a = 0.5, f_\alpha = 0.75, f_h = 0.75\) and \(\varphi = \pi/2\) and the Reynolds number is defined by \(Re = 2f_h \rho \pi h_a c/\mu\) because the free stream velocity is null. Average lift coefficients \(\bar{c}_l\) and \(\eta\) efficiencies were obtained simulating the hovering of a NACA 0012 airfoil, and are plotted as function of the Reynolds number in Figures 5a and 5b. The Reynolds number covers the range \(100 < Re < 1000\).

On the other hand the wake topology is studied in Fig. 6. It can be observed the LEV and TEV (trailing edge vortex) at \(Re = 150, \alpha_a = 20^\circ\) and \(\varphi = \pi/2\), caused by some stroke and its motions indicated by arrows. Three typical mechanism needed to understand the behavior of hovering are present: wake capture and diffusion effects, dynamic stall and roll-up effects.
Figure 5: Hovering motion over NACA 0012 airfoil with $h_a = 0.5$, $f_a = 0.75$, $f_h = 0.75$ and $\varphi = \pi/2$. (a) average lift coefficient $\bar{c}_l$. (b) efficiency $\eta$ hovering motion for two pitching amplitudes.

Figure 6: Velocity contours in hovering motion to NACA 0012 at $Re = 150$, $\alpha_a = 20^\circ$ and $\varphi = \pi/2$. (a) $t = 0.6s$, (b) $t = 0.92s$, (c) $t = 1.20s$, (d) $t = 1.64s$, (e) $t = 2.00s$, (f) $t = 2.26s$. (LEV_C and TEV_C are the previous stroke vortexes to capture and LEV_N the new vortex generated).

5 CONCLUSIONS AND FUTURE WORK

What has been sought with this work is the confirmation that current computational knowledge can to the ultra-low Reynolds number applications ($Re < 10000$), be extended. The main assumptions made about the flow field are: two dimensional incompressible, fully laminar steady and unsteady flows. The fully laminar assumption is the most physically accurate in the range of Reynolds numbers and angles of attack of interest here.
5.1 Steady Analysis

Two dimensional analysis allows a broad spectrum of parameters to be considered and provides a baseline for future more detailed studies. The geometric parameters investigated included thickness, camber etc, and it was intended to address the fundamental question of whether section geometry is still important at ultra-low Reynolds number. The most obvious effect of operating at ultra-low Reynolds numbers is a large increase in the section drag coefficient, however, the increase in drag is not reciprocated in lift, resulting in a large reduction in the $L/D$.

A ultra-low Reynolds numbers flow, is dominated by viscosity and the so called Boundary Layer concept is no longer applicable. Here it is generalized as the lower velocity viscous flow region adjacent to the body over which the pressure gradient normal to the surface is almost null. The extended constant pressure from the surface implies that the wing section effective geometry is significantly altered. As a result the pressure recovery is reduced and besides impacting on drag, at positive angles of attack the large effect is on lift. Viscous effects in thin wing sections thickness, significantly reduce the leading edge suction peak and the associated reduction in slope of the adverse pressure recovery, delays the onset of the stall. It can be stated that leading edge separation is delayed in thin wing sections and trailing edge separation is delayed in thicker sections. This behaviour could be potentially beneficial to lifting performance.

The effects of camber do not differ significantly from those observed at much higher Reynolds numbers. The fact that as the Reynolds numbers and section maximum thickness are reduced the details of the thickness distribution becomes less relevant, it allows to conclude that the camber-line is the dominant factor in performances.

5.2 Unsteady analysis

The highlight of the Finite Element software here used, is the ability to create mobile grids needed to simulate unsteady flights like heaving, flapping and hovering.

Symmetric wing sections are considered in studying the heaving motion. Average thrust coefficients and propulsion efficiencies are computed for given motion frequencies, and are plotted in terms of a Strouhal number determined using the amplitude of heaving. As a help for understanding the simulation results, figures are shown where velocity contours for two wing sections are compared at different times. The generation and displacement of vortices as the wing section executes the heaving motion, are well described.

Combinations of pitching and heaving motions (flapping) were simulated for symmetric wing sections, and thrust coefficients and propulsive efficiencies determined. Maximum pitch and vertical displacement amplitudes were taken as variables of plotting.

Hovering, is perhaps the type of flight that capture the greatest interest in the development of MAV. Average lift coefficients and power efficiencies are obtained simulating the flight of a symmetric wing section and the results plotted in terms of an ad-hoc Reynolds number defined taking into account that there is no free-stream. Classical aspects of
the hovering flight (leading and trailing edge already generated vortices, its capture and diffusion by the wake, and corresponding new generation), are shown.

5.3 Future area of research

So far, the fluid dynamic studies have been conducted and applied to a rigid 2D model. It is intended to extend first, the ultra-low Reynolds number area of research to 3D rigid finite span lifting wings, to the development of the viscous flow region adjacent to the surface, to describe stall patterns and wakes coming forth. Later, the rigid wing will be replaced by an elastic model and the coupling fluid-structure accounted for.

REFERENCES


CONTROLABLE PLANAR NOZZLES FOR MICRO THRUSTERS

P. H. PEDREIRA¹, J.R. LAURETTA¹ AND S. D'HERS²

¹Instituto Tecnológico de Buenos Aires
Departamento de Ingeniería Mecánica
Av. E. Madero 399, C1106ACD, Buenos Aires, Argentina.
popedreir@itba.edu.ar
rlaure@itba.edu.ar

²Instituto Tecnológico de Buenos Aires
Departamento de Ingeniería Mecánica - Centro de Mecánica Computacional
Av. E. Madero 399, C1106ACD, Buenos Aires, Argentina.
sdhers@itba.edu.ar

Key words: supersonic nozzles, compressible flow, CFD, micro thrusters.

Abstract. Planar nozzles can be used as component of a small propulsion system required for the orbital maneuvering of micro and nanosatellites. These nozzles could be machined using efficient design methods with optimum control of geometry and surface roughness, in contrast with axis-symmetrical nozzles where complex contours are difficult to machine.

It is known that throat active control is hard to achieve in axis-symmetrical geometry and in particular in the micro scale, however it can be done in planar nozzles. Hence to get a controllable nozzle at low cost, a planar nozzle was designed using the method of characteristics to get a hypersonic contour. Profile was then machined in AISI 1010 steel sheets using laser cutting techniques with a manufacturing cost less than 5% of a similar conical nozzle.

Tests were carried out in a vacuum chamber under several working conditions including throat area, measuring pressure, thrust, mass flow and temperature and correlating them to an ANSYS CFD model.

Both the CFD model and tests show that planar nozzles have no performance disadvantage over axis-symmetrical nozzles. Controllability analyses were performed assuming a feasible mechanism for orbital thrust control using an improved code based on the method of characteristics and testing. The designed system actuates on nozzle throat area and expansion ratio moving nozzle contours providing real-time thrust control. Tests showed that the thrust-inlet pressure ratio is a linear function of such displacement, and that no additional losses are introduced in the system. Hence it can be concluded that planar nozzles technology is suitable for small scale propulsion systems, with high impulse efficiency and controllable behavior.
LIST OF SYMBOLS

V: velocity modulus
r: throat radius
h: throat height
θ: tilt from nozzle axis
M: Mach
k: specific heat ratio
ṁ: mass flow
ST: throat section
P: absolute pressure
T: temperature
ρ: density
R: gas constant
TH: thrust
Δke: change in kinetic energy
E: expansion ratio

Subscripts
exit: conditions at nozzle exit
0: stagnation conditions
s: isentropic

1 INTRODUCTION

Miniaturization of satellites has led to the need of smaller propulsion systems [1]. As a result, micro nozzles are needed in order to provide the low levels of impulse required for the satellite pointing. Such nozzles are often axis-symmetrical and have to be machined using electrochemical etching [2] or other specialized techniques. This implies high costs and the need of expensive equipment. In addition, the impossibility of controlling nozzle expansion and throat diameter decreases the flexibility of the propulsion system and adds the need of auxiliary systems such as pressure regulators for controlling mass flow, as seen in the equation given below for a choked nozzle [3]

\[ \dot{m} \propto \frac{S_T P_0}{\sqrt{T_0}} \]  \hspace{1cm} (1)

A planar nozzle could overcome the stated issues, as they can be machined from very thin metal sheets using laser cutting techniques at a very low cost and then finishing the internal surfaces to the desired quality. Efficient design methods for the nozzle contour can be used without increasing machining costs [4].

A novel mechanism for thrust control is proposed. Nozzle thrust could be controlled by displacing nozzle sides, as shown in Figure 1. This modifies both nozzle throat area and expansion ratio in a simple fashion, providing thrust control. This variable configuration allows operation under different supply pressures saving the energy loss due to pressure...
regulation devices and heaters and reducing satellite components.

Figure 1: Nozzle profile view

2 NOZZLE DESIGN & EXPERIMENTAL SETUP

The nozzle contour was designed using the method of characteristics (MOC) [5] using a power series expansion of the Euler equations, centered in the nozzle throat [6], as boundary conditions as MOC requires supersonic boundary conditions. The contours obtained for different target Mach are shown in Figure 2. The boundary conditions are set in MOC calculations as to get an axial flow profile at the nozzle exit plane. As higher Mach gives lower pressures at the nozzle exit plane, Mach is limited to the working vacuum pressure of the vacuum chamber. For the present work, the profile for Mach 4.41 was selected.

Figure 2: MOC obtained profile for several Mach

The nozzle profile is machined in AISI 1010 steel sheets using a laser cutting machine. Figure 3 shows a scheme of the load cell mounting and the test bench.
Experiments are performed in a vacuum chamber. Vacuum is obtained using a two-stage, oil sealed vacuum pump, setting a pressure between 70 and 500 Pa for the experiments. The test bench is instrumented with vacuum and temperature sensors, a hot wire mass flow meter, stagnation temperature and pressure sensors and a load cell for thrust measurement. Air pressure is controlled with a manual pressure regulator. Tests are carried at desired outlet (vacuum) and inlet pressures opening a solenoid feed valve for a few seconds. This process is repeated for different nozzle throat opening, with varying pressures.

3 CFD MODELING

A 2D CFD model is developed to incorporate into the analysis phenomena that MOC cannot capture. The model was developed in CFX 14.5. This model solves the complete nozzle flow given the nozzle geometry and flow stagnation conditions. The propellant fluid is modelled using the ideal gas law with constant specific heat ratios:

\[ P = \rho RT \]  

The computational model is shown in Figure 4. At the open boundaries, opening boundary conditions are used, which allow flow in and out from the domain. Along the symmetry plane, symmetry boundary conditions are set. For turbulence, the Shear Stress Transport SST \( \kappa-\omega \) is adopted [7,8].

The free slip condition and adiabatic wall options are applied to nozzle walls, hence the boundary layer is not considered.
4 RESULTS

A nozzle contour cut in 2mm thick steel and 600 μm initial throat height is used. Thrust and mass flow are shown in Figure 5, compared to MOC theory and CFD results.

Figure 5: Thrust and mass flow for 2mm width nozzle

Good agreement between the different methods can be observed. As chamber pressure is set to match the design nozzle exit pressure, the thrust equation becomes:

\[ T_H = \dot{m}V_{exit} \]  

(3)

As the mass flow measures is very close to the theoretical mass flow calculation, it can be assumed that:

\[ V_{exit} \approx V_{\text{theoretical exit}} \]  

(4)

Isentropic efficiency, assuming adiabatic expansion of a nozzle is [9],

\[ \eta_{\text{nozzle}} = \frac{\Delta k_e}{\Delta k e_{\text{S}}} = \frac{V^2_{\text{exit}}}{V^2_{\text{exit,S}}} \]  

(5)
where the first quotient is the change in kinetic energy ratio. Experimental results show that isentropic efficiency is near unity. Thus, the expansion in the nozzle is quasi-isentropic. This means that the process is quasi-reversible. Hence it can be assumed that boundary layer effects are negligible for the prediction of thrust and mass flow in the proposed range. Thus properly designed planar nozzles have no disadvantages compared to axis-symmetrical ones.

Computational predicted design Mach differs in less than 0.2% from MOC value. In addition, CFD results show almost axial exit velocity distribution, as depicted in Figure 6a and 6b. This shows that nozzle contour effectively guides the flow field in the axial direction, reducing divergence losses.

**Figure 6a:** Axial Mach distribution.

**Figure 6b:** Deviation from axial direction in exit plane.
5 THRUST CONTROL

The nozzle control discussed actuates over nozzle throat, hence changing the expansion ratio and profile. Nozzle throat is of paramount importance as it regulates mass flow. The other parameters affect nozzle performance by changing exit Mach and causing the profile not to axialize the flow. As the profile cannot remain optimum when throat area changes, a Modified MOC model to address the growth of boundary layer and loss of axiality is implemented, named MMOC. It incorporates the concept of boundary layer displacement thickness [10]. The boundary layer displacement is calculated at the throat of the nozzle in order to fit mass flow measurements, for a given displacement. Then the nozzle profile is displaced to get a virtual contour. Usual MOC calculations are then performed with this virtual profile. Thrust is accurately predicted.

Results allow us to conclude that boundary layer effects in the nozzle can be regarded as a contour displacement that reduced the effective step of the nozzle, without introducing any losses in the system. Results are plotted in Figure 7a and 7b. The linear relation between inlet pressure and thrust is kept for every displacement tested.

![Figure 7: Variable nozzle. The number indicates throat height [mm] for different displacements.](image)

![Figure 7: Nozzle controllability is shown in the right side.](image)

Tests also show that the thrust-inlet pressure ratio is a linear function of the displacement,
as shown in Figure 7b. This is a key factor for controlling the nozzle in an actual system as only two functional parameters have to be specified. In this way, by measuring inlet conditions, the control system can adjust the nozzle step to achieve the desired thrust.

6 CONCLUSIONS

A planar nozzle capable of being actuated was designed using MOC, simulated using CFD and tested in a test bench. The whole development showed that:

- Planar nozzles match the performance of axis-symmetrical nozzles, with lower manufacturing costs.
- Miniaturization can be achieved without the loss of performance. This was proved for a minimum nozzle thickness of 2mm.
- Variable thrust with constant inlet pressure can be obtained from the planar nozzles in a simple fashion for real time thrust controls. The obtained thrust vs. pressure relation varies linearly with the displacement.

Then it can be concluded that planar nozzles technology is suitable for small scale propulsion systems, with high impulse efficiency and controllable behavior.

ACKNOWLEDGEMENTS

The authors would like to thank A. Beceyro Ferrán and L. De Cunto for their collaboration in the experimental work, to Satellogic for their generous donation of the test bench and to Instituto Tecnológico de Buenos Aires for the funding granted through ITBACyT.

REFERENCES


APPLICATION OF PARTIALLY-AVERAGED NAVIER-STOKES SIMULATIONS FOR BLUFF BODY AERODYNAMICS

M. MIRZAEI¹, S. KRAJNOVIĆ²

Department of Applied Mechanics, Division of Fluid Mechanics, 41259 Gothenburg, Sweden
¹ marmir@chalmers.se
² sinisa@chalmers.se

Key words: Partially-Averaged Navier-Stokes, PANS, LES, Ahmed body.

Abstract. This paper presents a study about the prediction of flows around simplified vehicles using Partially-Averaged Navier-Stokes (PANS). Numerical simulations are performed for two simplified vehicles: a slanted-back Ahmed body at Re=30 000 and a squared-back Ahmed body at Re=300 000. A comparison of the resolved and modeled physical flow scales is made with corresponding LES and experimental data for a better understanding of the performance of the PANS model. The PANS model is compared for coarse and fine grid resolutions and it is indicated that even a coarse-grid PANS simulation is able to produce fairly close flow predictions to those from a well-resolved LES simulation.

1 INTRODUCTION

The standard Reynolds Averaged Navier-Stokes (RANS) approach models all the flow scales and is not able to adequately predict the highly turbulent wake region and flow structures of bluff body flows. Large eddy simulations (LES), on the other hand, is superior to RANS methods in strongly separated flows. LES is however a computationally demanding technique, and it is not feasible to compute complex vehicle flows at higher Reynolds numbers with LES. The LES simulations are thus always performed for the flows with Reynolds numbers that are several times lower than the operational ones.

The Partially-Averaged Navier-Stokes (PANS) approach is a method originally proposed by Girimaji et al. (2003) (1). Girimaji (1) suggests that the RANS closure coefficients are formally valid only when the averaging is performed over all fluctuating scales of motion. In order to resolve fluctuating scales, the RANS model coefficients must be suitably modified, according to the desired resolution, in a physically correct manner. The PANS model changes seamlessly from RANS to the direct numerical solution (DNS) of the Navier-Stokes equations and is purported to provide the best possible physical fidelity.
on any given numerical grid. Lakshmipathy and Girimaji (2) used PANS based on the $k-\varepsilon$ turbulence model to simulate the flow past a circular cylinder. They found that, even though PANS could highly improve the separation point prediction compared to RANS, it still suffers from the late separation prediction in flow around two-dimensional and three-dimensional cylinders.

Basara et al. (3) proposed a new PANS $k-\varepsilon-\zeta-f$ model that, highly enhances near-wall flow predictions. The PANS $k-\varepsilon-\zeta-f$ model was used for several geometries, and significant improvements were observed as compared to the original $k-\varepsilon$ PANS model. Basara et al. (6) evaluated the proposed PANS model by simulating the flow around a three-dimensional circular cylinder mounted vertically on a flat plate. Their results indicated that the four-equation PANS model based on $k-\varepsilon-\zeta-f$ reproduces nearly all of the unsteady flow features around the cylinder, in good agreement with experiments and LES. The aim of the present study is to investigate the capability of the PANS $k-\varepsilon-\zeta-f$ model to predict the flow around two different Ahmed bodies. The Ahmed body is one of the most popular models employed as a simplified vehicle shape to predict the fundamental flow physics associated with vehicles’ wake. The original shape of the body and flow structure around it can be found in (4).

The organization of the paper is as follows. Sections 2 and 3 present a brief summary of the PANS $k-\varepsilon-\zeta-f$ and the LES models, corresponding equations and boundary conditions. The results are described in section 4 and are presented separately for a slanted-back Ahmed body at a low Reynolds number and a squared-back Ahmed body at a high Reynolds number. To verify the accuracy of the predictions, the results are compared with the LES results of the present study and the available LES and PIV data (5).

2 NUMERICAL MODELING METHOD

2.1 PARTIALLY-AVERAGED NAVIER-STOKES (PANS)

The PANS equations read (1):

$$\frac{\partial U_i}{\partial t} + U_j \frac{\partial U_i}{\partial x_j} + \frac{\partial \tau(V_i, V_j)}{\partial x_j} = -\frac{1}{\rho} \frac{\partial p}{\partial x_i} + \nu \frac{\partial^2 U_i}{\partial x_j \partial x_j}$$

where $V_i$ indicates the instantaneous turbulent velocity field and is decomposed in two parts as $V_i = U_i + u_i$. To close the system of the partially-averaged Navier-Stokes equations, Girimaji et al. (1) proposed using the conventional eddy viscosity concept, which reads as $\tau(V_i, V_j) = -2\nu_u S_{ij} + 2/3 k_u \delta_{ij}$, where $S_{ij} = 1/2(\frac{\partial U_i}{\partial x_j} + \frac{\partial U_j}{\partial x_i})$ is the resolved strain-tensor and $\nu_u = c^2 \frac{k^2}{\varepsilon}$ is the PANS eddy viscosity.

To formulate the PANS model eddy viscosity, Girimaji (1) employed the standard $k-\varepsilon$ model to derive two transport equations for the unresolved kinetic energy ($k_u$) and
unresolved dissipation ($\varepsilon_u$):

$$\frac{\partial k_u}{\partial t} + U_j \frac{\partial k_u}{\partial x_j} = P_u - \varepsilon_u + \frac{\partial}{\partial x_j} \left( \nu_u \frac{\partial k_u}{\partial x_j} \right)$$

(2)

$$\frac{\partial \varepsilon_u}{\partial t} + U_j \frac{\partial \varepsilon_u}{\partial x_j} = C_{\varepsilon_1} P_u \frac{\varepsilon_u}{k_u} - C_{\varepsilon_2} \varepsilon_u^2 + \frac{\partial}{\partial x_j} \left( \nu_u \frac{\partial \varepsilon_u}{\partial x_j} \right)$$

where the model coefficients are given as $C_{\varepsilon_2} = C_{\varepsilon_1} + \frac{f_k}{f_c} (C_{\varepsilon_2} - C_{\varepsilon_1})$, $\sigma_{k_u} = \sigma_k \frac{k^2}{f_c}$ and $\sigma_{\varepsilon_u} = \sigma_{\varepsilon} \frac{k^2}{f_c}$.

In the derivation of the transport equations for $k_u$ and $\varepsilon_u$, two parameters, $f_k = k_u/k$ and $f_c = \varepsilon_u/\varepsilon$, were introduced that relate to the unresolved to total ratios of kinetic energy and dissipation, respectively. The unresolved to total kinetic energy ratio, $f_k$, was originally prescribed as a constant (1). Basara et al. (6) modified the $f_k$ parameter as a variable that depends on the grid spacing as follows:

$$f_k \geq \frac{1}{\sqrt{c_\mu}} \left( \frac{\Delta}{\Lambda} \right)^{2/3}$$

$$\Lambda = \frac{k^{3/2}}{\varepsilon}$$

(3)

(4)

where $\Delta$ is the geometric averaged grid cell dimension ($\Delta = (\Delta_x \Delta_y \Delta_z)^{1/3}$) and $\Lambda$ is the integral length scale of turbulence. Basara et al. (3) later developed a new four-equation PANS model based on the $k - \varepsilon - \zeta - f$ RANS model, called as PANS-$k - \varepsilon - \zeta - f$, that could more adequately predict the complex behavior of boundary layer separations around smooth bluff-bodies. The $k - \varepsilon - \zeta - f$ model is derived from the $v^2 - f$ model in which a transport equation for $\zeta$ ($\zeta = v^2/k$, where $k$ is the turbulent kinetic energy) was employed rather than $v^2$. Furthermore, the $f$ equation, which is related to the pressure strain term, is also computed, and this improved the near wall predictions. The PANS-$k - \varepsilon - \zeta - f$ model equations are briefly presented below (a more detailed derivation can be found in (3)):

$$\frac{\partial k_u}{\partial t} + U_j \frac{\partial k_u}{\partial x_j} = P_u - \varepsilon_u + \frac{\partial}{\partial x_j} \left[ (\nu + \frac{\nu_u}{\sigma_{k_u}}) \frac{\partial k_u}{\partial x_j} \right]$$

$$\frac{\partial \varepsilon_u}{\partial t} + U_j \frac{\partial \varepsilon_u}{\partial x_j} = C_{\varepsilon_1} P_u \frac{\varepsilon_u}{k_u} - C_{\varepsilon_2} \varepsilon_u^2 + \frac{\partial}{\partial x_j} \left[ (\nu + \frac{\nu_u}{\sigma_{\varepsilon_u}}) \frac{\partial \varepsilon_u}{\partial x_j} \right]$$

$$\frac{\partial \zeta_u}{\partial t} + U_j \frac{\partial \zeta_u}{\partial x_j} = f_u - \frac{\zeta_u}{k_u} P_u - \frac{\zeta_u}{k_u} \varepsilon_u (1 - f_k) + \frac{\partial}{\partial x_j} \left[ (\nu + \frac{\nu_u}{\sigma_{\zeta_u}}) \frac{\partial \zeta_u}{\partial x_j} \right]$$

$$L_u^2 \nabla^2 f_u - f_u = \frac{1}{T_u} (c_1 + c_2 \frac{P_u}{\varepsilon_u}) (\zeta_u - \frac{2}{3})$$

$$\nu_u = \frac{C_\mu}{k_u} \frac{k^2_u}{\varepsilon_u}$$
where \( C_\ast^* = C_{\varepsilon 1} + f_k(C_{\varepsilon 2} - C_{\varepsilon 1}), C_{\varepsilon 1} = 1.4(1 + 0.045/\sqrt{\varepsilon}), C_\mu = 0.22, c_1 = 0.4, c_2 = 0.65 \) and \( C_{\varepsilon 2} = 1.9 \). \( L_u \) is the length scale and \( T_u \) is the time scale of the PANS model, defined respectively as:

\[
T_u = \max[k_u^3/\varepsilon, C_\tau(\nu^3/\varepsilon)^{1/2}]
\]

\[
L_u = C_L \max[k_u^3/\varepsilon, C_\eta(\nu^3/\varepsilon)^{1/4}]
\]

where \( C_L \) and \( C_\tau \) are equal to 0.36 and 6.0, respectively. It is assumed that the numerical meshes support the cut-off in the energy containing scales and inertial range, meaning that \( \varepsilon_u = \varepsilon \), thus \( f_\varepsilon = 1 \) (3).

2.2 LARGE EDDY SIMULATION (LES)

The governing LES equations employed in the present study are the incompressible Navier-Stokes and the continuity equations. The algebraic eddy viscosity model proposed by Smagorinsky (7) was used in this study. The Smagorinsky model represents the anisotropic part of the SGS stress tensor \( \tau_{ij} \) as \( \tau_{ij} = -\nu_{sgs} \bar{S}_{ij} \), where \( \nu_{sgs} \) is defined as \( \nu_{sgs} = (C_S f \Delta)^2(\bar{S}_{ij} \bar{S}_{ij})^{1/2} \). \( C_S = 0.1, f \) is the van Driest damping function and is defined as \( f = 1 - \exp(-n^+ / 25) \), where \( n \) is the wall normal distance. \( \Delta \) is the filter width and is defined as \( \Delta = (\Delta_1 \Delta_2 \Delta_3)^{1/3} \), where \( \Delta_i \) are the computational cell sizes in three coordinate directions.

2.3 BOUNDARY CONDITIONS AND NUMERICAL METHODS

The AVL FIRE commercial finite volume solver is employed to solve both PANS and LES equations. The SIMPLE algorithm is used to update the pressure and velocity fields. The second-order three-time level scheme is used for time discretization. Convective fluxes in the LES model are approximated by a blend of 99% central differences of second-order accuracy and of 1% upwind differences. For the PANS model, the convective fluxes are estimated by the AVL SMART method (3), which is a second-order bounded scheme.

3 RESULTS: PANS vs. LES FOR SLANTED BACK BODY

The PANS-\( k - \varepsilon - \zeta - f \) model is used here for the flow around a three-dimensional slanted-back Ahmed body with a slant angle of 25°. It should be noted that the case of the slant angle of 25° has been recognized to be a critical angle, for which all RANS models fail to correctly predict the flow.

3.1 COMPUTATIONAL DOMAIN AND GRID SPECIFICATIONS

Figure 1 shows a schematic view of a slanted-back Ahmed body placed in the computational domain. The Reynolds number based on the model height is \( Re_H = 30\,000 \). The dimensions of the model are: height \( H = 297 \) mm, length \( L = 3.7H \), width \( W = 1.18H \).
and the ground clearance is \( h = 0.17H \). The back light angle is \( \alpha = 25^\circ \). A uniform velocity profile of \( U_{in} = 1.53 \text{ m/s} \) is used at the inlet. The homogeneous Neuman boundary condition is used at the outlet. The lateral sides and the ceiling of the computational domain are treated as slip surfaces. The no-slip boundary condition is used on the surface of the body and channel floor. The channel floor is set as a moving ground having the same velocity as the inlet boundary condition. Simulations are performed using two computational grids consisting of 2.5 and 5 million cells, respectively. For both coarse and fine grids, the averaged wall normal resolution in the model is \( n^+ = 0.7 \). The grid resolution in the streamwise direction and spanwise direction are \( \Delta s^+ < 100 \) and \( \Delta l^+ < 30 \) and \( \Delta s^+ < 150 \) and \( \Delta l^+ < 40 \) for the coarse and fine PANS simulations, respectively. The PANS results are compared with a well-resolved LES simulation of the present study with 9 million computational cells. For the LES grid, the resolution in the streamwise and spanwise directions are \( \Delta s^+ < 70 \) and \( \Delta l^+ < 20 \), respectively. Here, \( n^+ = nu_\tau/\nu \), \( \Delta s^+ = \Delta su_\tau/\nu \) and \( \Delta l^+ = \Delta lu_\tau/\nu \), where \( u_\tau \) is the wall friction velocity. The time step, \( dt^* = dtU_\infty/H \), for all the simulations is selected to be \( 3 \times 10^{-3} \), leading to the CFL number having a value of less than one in over 99% of the cells.

3.2 COMPARISON OF FLOW STRUCTURES

Figure 2 presents the time-averaged streamlines and streamwise velocity component contours in the symmetric plane behind the body, predicted by PANS and LES simulations. The PANS results are shown for both coarse and fine grids. As can be observed, the reattachment of the separated flow on the top edge of the slanted surface is in a good agreement with the prediction of LES in both the coarse- and fine-grid PANS simulations. However, the length of the separated bubble obtained in the coarse-grid PANS is overpredicted, compared to LES, and covers nearly the whole back surface, while the length is shorter for the fine-grid. The predictions of the thickness of the bubble in the wall-normal direction and the center of that are close in the coarse-grid PANS and LES, whereas in the fine-grid PANS the bubble is narrower and moves a little toward the top edge. The size and the position of the upper and lower vortices in the wake region of the body given by the coarse and fine PANS are also in a good agreement with LES.

Figure 3 compares the time-averaged streamwise \( (u) \) velocity component, obtained by PANS with coarse and fine grids and LES along three different lines in the wake of the body. The general observation is that both coarse- and fine-grid PANS simulations predict trends similar to those provided by LES. The reversed-velocity region along the first line predicted by the coarse-PANS grid is in good agreement to the LES one, while the reversed velocity is smaller for the fine-grid PANS due to the smaller size of bubble over the rear surface. A large discrepancy is observed between the coarse-grid PANS and LES along the second line, while a considerably better prediction is produced by the fine-grid PANS.

In order to visualize the level to which the flow structures are resolved or modeled, the instantaneous contours of \( f_k \) for both coarse- and fine-grids PANS simulations are pre-
sented in figure 4. The results are shown for input $f_k$, defined in Eq. 3, as well as the output value of $f_k$ resulting from the unresolved to total turbulent kinetic energy ratio, $f_k = k_u/k_{tot}$. As can be observed in both the coarse and fine grids, the $f_k$ values are high in the upstream of the body and low around the body and in the wake region. A higher level of flow structures are resolved by refining the grid resolution, resulting in a smaller values of unresolved kinetic energy, $k_u$, and finally, smaller values of $f_k$. As required by the model, the values of output $f_k$ are lower than the input $f_k$, proving that the PANS model adequately resolves the flow structures.

A more evident indication of the performance of the PANS model can be illustrated by plotting the instantaneous iso-surfaces of the second invariant of the velocity gradient ($Q = -1/2(\partial u_i/\partial x_j \partial u_j/\partial x_i)$). Figure 5 visualizes the instantaneous iso-surfaces of Q in the near-wake region obtained by the PANS and LES simulations. The iso-surfaces are colored by the streamwise component of averaged vorticity ($\omega_x$). Note that the velocity component, $u_i$, represents the turbulent fluctuation in DNS or LES, whereas it is the partially-averaged velocity in PANS. Therefore, as was discussed in (8), it is important to explain the flow structures properly based on different methods. Figure 5 (a) indicates that LES could resolve the flow structures in the near wake region owing to the well-resolved grid resolution around the body. $f_k$ contours for coarse-grid PANS (figure 4 (a) and (b)) show that more portions of the flow scales close to the near-wake region are modeled and less of them are resolved. Refining the grid decreases the ratio of the unresolved to total kinetic energy, $f_k$ (see figure 4) and, thus, more instantaneous flow structures are released in the wake region (figure 5 (c)). It is worth noting that both the coarse- and fine-grid PANS are able to produce the C-pillar trailing vortices along the lateral edges. The fine-grid PANS includes a better grid resolution in the wake region and captures a higher level of the flow structure. Therefore, the trailing vortex propagates further downstream and provides results closer to LES as compared to the coarse PANS results.

4 RESULTS: PANS vs. LES FOR SQUARED BACK AHMED BODY

In this section, the PANS-$k-\varepsilon-\zeta-f$ model is employed to simulate a squared-back Ahmed body. An identical squared-back model with the same flow conditions and Reynolds number was used in the numerical and experimental investigations using LES and PIV, respectively (5).

4.1 COMPUTATIONAL DOMIAN AND GRID CHARACTERISTICS

The simplified vehicle model is a squared-back Ahmed body. A schematic view of the model is shown in figure 6. The body is located on four cylindrical supports with an oval cross-section and a ground clearance of 0.168H (see figure 6). The Reynolds number based on the height of the model and the inlet velocity is $Re_H = 300,000$. As in the LES simulations, the flow parameters are averaged over $t^* = tU_\infty/H = 500$ time units. The numerical model and boundary conditions are the same as those used in the previous sec-
tion. However, in order to match the set-ups used in the LES and PIV measurements (5), the slip boundary condition is used in the first part of the ground and a no-slip boundary condition is employed for the rest of it. Two coarse and fine grid resolutions with 8 and 12.5 million cells, respectively, are used for the PANS simulations. The computational grid employed in the LES simulations of (5) included 34 million cells, which is considerably finer than the PANS grids in the present study.

4.2 COMPARISON OF FLOW STRUCTURES

The side-view of the time-averaged streamlines and streamwise velocity component ($u$) contours in the symmetry planes of the wake region are shown in figures 7. The PANS results are presented for the fine grid and are compared with the corresponding results of LES and PIV (5). Two vortical structures, one over the other, can be observed in the averaged flow structure in figure 7, where the upper vortex is larger than the lower one. The size and the location of the center of the lower vortex are well-predicted. The PANS model provides a prediction closer to PIV in terms of the length of the wake behind the body, compared to the LES one. However, the size of the upper vortex is smaller than that provided by LES and PIV. The position of the upper vortex is further downstream compared to the LES and PIV results.

The averaged streamwise velocity component ($u$) profiles along three different lines in the symmetry plane of the body are shown in figure 8. The selected lines are located in the streamwise distances $0.17H$, $0.5H$ and $0.84H$ with respect to the base wall of the body. The velocity profiles are presented for both grids to show the improvement of the results between the coarse and fine grids. The results are also compared with the corresponding results of LES and PIV (5). The general variation of the profiles indicates that the PANS model produces quite similar results to those of LES and PIV. Owing to the mismatch of the upper vortex, some differences are observed around the center of the wake. Refinement of the grid in the center of the wake can reasonably improve the agreement of the velocity profiles, however, and this indicates that there is hope in employing a higher resolution grid in this region to provide predictions even closer to the LES and PIV results.

Figure 10 presents the top-views of the instantaneous contours of $f_k$ for the coarse- and fine-grid PANS simulations. The plots are shown for the input value of $f_k$ (Eq. 3). Similar to the results of the slanted body, the $f_k$ values are high at the upstream of the body and are low around the body and in the wake region. A higher level of flow structures are resolved by employing a finer grid resolution, which results in smaller values of the input $f_k$.

Figure 9 provides a comparison between the resolved flow structures obtained by the coarse and fine grids. The fine-grid PANS simulation captures a significantly higher portion of the resolved flow structures close to the walls as well as the wake of the body, compared to the coarser one. The flow structures in this figure and averaged velocity profiles in figure 8 predicted by coarse and fine grids show that, despite the good agreement
between the velocity profiles from coarse and fine grids, there is a large difference between the flow structures predicted by these two grids. Figure 11 plots the profiles of normalized modeled and resolved shear stress \(\frac{u'v'\Delta}{\frac{1}{2}U_{in}^2}\) along four lines on the symmetry plane of the wake. The locations of the lines are the same as those presented in figure 8. The results are shown for the fine grid PANS simulation. As can be seen, the modeled values are significantly smaller than the resolved ones along all four lines, which indicates the ability of the PANS model to resolve the scales to a high level in the near-wake region.

5 CONCLUSIONS

The paper presented an evaluation of the Partially-Averaged Navier-Stokes (PANS) method based on the \(k-\varepsilon-\zeta-f\) RANS model for simplified vehicle flows. The simulations were made for a slanted-back Ahmed body at Re=30,000 and a squared-back Ahmed body at Re=300,000. The PANS results were compared with the corresponding LES simulation and previous LES and PIV data provided in (5), respectively. A comprehensive study of the resolved and modeled physical flow structures was presented for both geometries and showed that the PANS model resolves the flow scales to a reasonably high level, compared to LES and experimental results. The simulations were performed for the coarse- and fine-grid resolutions and, it was recognized that even a coarse-grid PANS simulation is able to produce fairly acceptable predictions. Some differences were observed between the wake structure predicted in the present study and those predicted by LES and PIV for the squared-back body. However, even the fine-grid resolution employed for the PANS simulations is significantly coarser than that in LES (12.5 million cells in PANS and 34 million cells in LES).

ACKNOWLEDGMENT

This work is supported financially by the Safer, Vehicle and Traffic Safety Centre at Chalmers. Software licenses were provided by AVL List GMBH. Computations were performed at the National Supercomputer Center (NSC) at LiU.

References


**Figure 1**: Schematic view of the computational domain and the slanted-back Ahmed body.

**Figure 2**: Averaged streamlines and streamwise velocity ($u$) contours in the symmetric plane for LES, coarse-grid PANS and fine-grid PANS.
Figure 3: Averaged streamwise velocity \( u \) profiles at three different selected locations.

Figure 4: Comparison of \( f_k \) computed based on the length scale (input \( f_k \), (a) and (b)) and unresolved to total kinetic energy (output \( f_k \), (c) and (d)) around the body from PANS with coarse and fine grids.

Figure 5: Iso-surfaces of instantaneous \( Q = 400 \) \((1/s^2)\) colored by streamwise vorticity component \( \omega_x \) predicted by (a) LES, (b) PANS with coarse grid and (c) PANS with fine grid.
Figure 6: A schematic view of the squared back Ahmed body.

Figure 7: Side view of the time-averaged streamwise velocity ($u$) contour and streamlines in the symmetric plane from PANS, LES (5) and experimental PIV (5).

Figure 8: Comparison of the averaged streamwise velocity ($u$) profiles predicted by PANS, LES (5) and experimental PIV data (5).
Figure 9: Iso-surfaces of instantaneous $Q = 5000$ (1/s$^2$) for PANS with coarse and fine grids colored by streamwise vorticity ($\omega_x$).

Figure 10: Comparison of the inlet values of $f_k$ for the coarse and fine grids.

Figure 11: Resolved and modeled turbulent shear stresses for the fine grid. Solid lines are resolved and dashed lines are modeled stresses.
STRUCTURAL MODELS AND MECHANICAL TESTS IN THE
DEVELOPMENT OF A COMMUNICATIONS SPACECRAFT

SANTIAGO TEMPONE*1, MARCOS FRANCESCHINI**
and EDUARDO N. DVORKIN***

*INVAP S.E.
A. Cmte. Luis Piedrabuena 4950, R4803CPV, Bariloche, Rio Negro, Argentina
stempone@invap.com.ar http://www.invap.com.ar/

** INVAP S.E.
A. Cmte. Luis Piedrabuena 4950, R4803CPV, Bariloche, Rio Negro, Argentina
mfranceschini@invap.com.ar http://www.invap.com.ar/

*** SIM&TEC S.A.
Av. Pueyrredón 2130 5to.A, C1119ACR, Buenos Aires, Argentina
edvorkin@simytec.com http://www.simytec.com

Key words: spacecraft, finite elements.

Abstract. The objective of this paper is to discuss the development of the computational structural
models that were used to check the structural strength and orbital alignment of the communications
spacecraft ARSAT-1, designed and built by INVAP.

1 INTRODUCTION

The objective of this paper is to discuss the computational structural models that were developed to
check the strength and orbital alignment of the communications spacecraft ARSAT-1, designed and
built by INVAP S.E. at Bariloche, Argentina under ARSAT S.A. requirements.

The results of the computational models were used, by the satellite developers, to check the
structural integrity of the spacecraft structure and its ability to withstand the in-orbit thermal cycles
without losing the communication devices pointing. In more details, it can be assessed that the output
of the computational models was used by the spacecraft developers for:

• verifying, at the structural design stage, that the structure had the required strength and
stiffness and that its dynamic behavior was compatible with the dynamic environment of the launching
vehicle and no undesirable resonances were to be encountered;

• planning the physical tests and establishing their limits in order to protect the integrity of the
spacecraft, of its instrumentation and of the testing machines.

After the execution of the physical tests the simulation results and experimental determinations
were confronted to validate the computational – experimental spacecraft qualification process.

Two consecutive physical models are normally used in the qualification of a spacecraft [1] [2]:

---

1 The work presented in this paper was developed while at SIM&TEC S.A.
• first the structural test model (STM): it is a full size model in which all the structural components are assembled but the actual spacecraft equipment is replaced by syntonized dummies. The wiring and insulation blankets are not installed on this model;

• finally the protoflight test model (PFM): it is basically the actual spacecraft in its launching configuration. Due to the presence of the wiring and insulation blankets this physical model presents more damping under dynamic excitations than the structural one.

In this paper some of the tests and numerical models used to qualify the STM are discussed.

The physical tests are:

• dynamic tests: sinusoidal frequency dependent excitations (sine sweep test and acoustic excitations). These tests simulate the mechanical excitations of the spacecraft inside the launcher;

• vacuum thermal loading test (TVAC): used to simulate the thermal cycles that undergoes the spacecraft in orbit.

While the objective of the dynamic tests is to check the structural integrity of the spacecraft when inside the launching vehicle, the objective of the thermal test is to check the structural integrity and alignment preservation of the spacecraft in orbit under thermal cycles.

The spacecraft structure is assembled using threaded connections; hence, bolted joints are very important structural components; therefore, special attention was given to the simulation of their behavior: an ad hoc elastic-plastic model was used to incorporate into the model the possibility of the frictional joints slipping and the effect, on the communication devices pointing, of the slipping hysteresis that may be accumulated during the planned length of the spacecraft mission.

The effect of the shocks induced by the pyrotechnic devices used for the separation of the spacecraft from the launch vehicle and for the deployment of the solar panels and antennas were experimentally investigated using the STM and the PFM.

All the possible failure modes that can be encountered during the testing, launching and orbiting were investigated with the developed computational models and mechanical tests.

In the second section of this paper the main features of the spacecraft finite element model are discussed.

In the third section the determination of natural frequencies is discussed while in the fourth section the numerical simulation of quasi-static tests (sine burst) and the sine sweep tests are also discussed.

In the fifth section the experimental-numerical procedure used for analysing the results of the acoustic tests are presented.

Finally the sixth section deals with the thermo-elastic model and the threaded connection model.

The finite element analyses were conducted using the commercial code FEMAP with NX Nastran.

2 MAIN FEATURES OF THE SPACECRAFT FINITE ELEMENT MODEL

2.1 Discretization

In Figure 1 the finite element model of the STM is displayed.
The spacecraft panels are constructed using sandwich plates with an aluminum honeycomb core and facings made with aluminum in some cases and with carbon fiber laminates in other cases. In both cases the model is developed using laminate plate elements. In this kind of elements, in which the stiffness of the core is very different from the stiffness of the facings, a proper choice of the shear correction factors is of utmost importance to produce realistic results [3].

The function of the spacecraft handling adapter (SHA) is to fix the spacecraft to the shaker test table. The configuration of the shaker adapter is the same for Slip Table (SL) (X and Y horizontal directions) and Head Expander (HE) (Z vertical direction).

The SL adapter is composed by one intermediate plate, one base plate, one superior ring (SHA) and eight force-links as one of the options to be analyzed and sixteen force-links for the second option to be analyzed, as shown in Figure 2.

The SHA is modeled using solid tetrahedral elements, while the plates are modeled with plate elements. Details of the mesh are displayed in Figure 3 and Figure 4.
The bolted connections are modeled using rigid links with a 6 d.o.f. spring elements (CBUSH elements) in series with it (Figure 5).

The properties of CBUSH elements corresponding to bolted joints on the adapter were adjusted considering its stiffness. This adopted value was obtained through to a comparison between numerical...
results from frequency analyses and results of the natural frequency test of the model performed at INVAP.

In Figure 6 the HE model is depicted.

Figure 6. HE finite element model

In Figure 7 it is shown a force-link and a scheme of its model representation; the total mass of each force-link is simulated using two mass elements at the same geometrical location but using two different nodes: linked to the upper and lower rigid links respectively; a CBUSH, used to simulate the force-link manufacturer specified stiffness, links both masses.

Figure 7. Force-link and its finite element model

The assembly of the SHA on the HE is shown in Figure 8.

Figure 8. Assembly SHA and HE
The total number of elements used in the different models is indicated in Table 1.

**Table 1. Elements used in the developed models**

<table>
<thead>
<tr>
<th>Element Type</th>
<th>ARSAT-1 STM # Elements</th>
<th>8 F-L SHA+SL # Elements</th>
<th>16 F-L SHA+SL # Elements</th>
<th>8 F-L SHA+HE # Elements</th>
<th>16 F-L SHA+HE # Elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bar</td>
<td>1,243</td>
<td>1,243</td>
<td>1,243</td>
<td>1,243</td>
<td>1,243</td>
</tr>
<tr>
<td>Beam</td>
<td>1,146</td>
<td>1,146</td>
<td>1,146</td>
<td>1,146</td>
<td>1,146</td>
</tr>
<tr>
<td>CBUSH</td>
<td>3,284</td>
<td>3,540</td>
<td>3,612</td>
<td>3,540</td>
<td>3,612</td>
</tr>
<tr>
<td>DOF spring</td>
<td>4,911</td>
<td>4,911</td>
<td>4,911</td>
<td>4,911</td>
<td>4,911</td>
</tr>
<tr>
<td>Plate</td>
<td>138,218</td>
<td>265,334</td>
<td>265,334</td>
<td>385,167</td>
<td>385,167</td>
</tr>
<tr>
<td>Laminate</td>
<td>276,682</td>
<td>276,682</td>
<td>276,682</td>
<td>276,682</td>
<td>276,682</td>
</tr>
<tr>
<td>Solid</td>
<td>0</td>
<td>206,488</td>
<td>206,488</td>
<td>206,488</td>
<td>206,488</td>
</tr>
<tr>
<td>Mass</td>
<td>355</td>
<td>371</td>
<td>387</td>
<td>371</td>
<td>387</td>
</tr>
<tr>
<td>Rigid</td>
<td>9,778</td>
<td>11,326</td>
<td>11,374</td>
<td>11,395</td>
<td>11,443</td>
</tr>
<tr>
<td>TOTAL</td>
<td>435,617</td>
<td>771,041</td>
<td>771,177</td>
<td>890,943</td>
<td>891,079</td>
</tr>
</tbody>
</table>

**2.2 Boundary conditions**

For modeling the “Hard Mounted” condition all the 6 degrees of freedom of the nodes shown in Figure 9 were fixed.

![Figure 9. STM Hard Mounted boundary conditions](image)

The SHA is fixed to the ST using 184 bolts as shown in Figure 10.
Figure 10. SHA fixation to the SL

The force-links (Figure 7) are fixed to the SHA and base plate as shown in Figure 11.

Figure 11. Force-links fixation bolts

When the HE is used, the fixations are shown in Figure 12.

Figure 12. Bolts used in HE configuration
3 NATURAL FREQUENCIES

With the developed model, the natural frequencies in the interval [0, 150 Hz], of the hard mounted STM were determined, totalizing a number of 171 natural modes in the prescribed interval, Figure 13.

The first flexural modes and the axial mode are listed in Table 2.

Table 2. First natural frequencies for Hard Mounted condition

<table>
<thead>
<tr>
<th>Mode</th>
<th>F (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>First bending mode in the y-direction</td>
<td>18.74</td>
</tr>
<tr>
<td>First bending mode in the x-direction</td>
<td>20.13</td>
</tr>
<tr>
<td>First axial mode</td>
<td>61.37</td>
</tr>
</tbody>
</table>

The STM is mounted on the shaker using the SHA, the HE and the force-links. One of the objectives in the design of those intermediate parts is to minimize the difference between the natural frequencies of the STM as assembled on the shaker (Flexible configurations) and the STM Hard Mounted natural frequencies.
As it can be seen comparing Figure 14 and Figure 15 the change in natural frequencies introduced by the force-links is quite small.

In Figure 16 and Figure 17, configurations corresponding to selected natural modes for the case of 8 force-links are shown for the STM assembled with the SL or HE.
4 MODELING OF THE SINE BURST AND SINE SWEEP TESTS

4.1 The sine burst test

The sine burst test is a test, in which the ground acceleration is applied with the function [2],
\[ g_{sb} = f(t) A_{sb} \sin(2\pi f_{sb}) \] (1)

where \( f(t) \) reaches its unit value after a number of cycles, remains equal to one from 5 to 10 cycles and then it descends to zero after a number of cycles. The frequency is \( f_{sb} < \frac{f_n}{3} \), where \( f_n \) is the lowest natural frequency; hence, this test is modeled as a static test.

The accelerations considered in the static analysis are selected to induce the defined qualification loads at the separation plane.

Therefore the following three loading case accelerations are determined:

<table>
<thead>
<tr>
<th>Load case</th>
<th>X-acc [g]</th>
<th>Y-acc [g]</th>
<th>z-acc [g]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.778</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>2.778</td>
<td>-1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>-8.49</td>
</tr>
</tbody>
</table>

The margins of safety in panels and bolts are determined using the mass loadings defined by Table 3.

4.2 The sine sweep test

A harmonic frequency response analysis is performed for 8 and 16 force-link configurations considering the adapter plus the STM plus the Slip Table / Head Expander. For each configuration, a
sine sweep in X, Y and Z direction is performed for the interval [0-100Hz]. An acceleration of 1g is imposed in nodes that represent the Slip Table and Head Expander fixations to test machine.

Two structural damping coefficients are used: 2.5% for 0 to 50Hz and 1% for 51 to 100Hz.

The analyses are developed considering 8 and 16 force-links; to illustrate on the results in Figure 18 the results corresponding to the 8 force-links case presented.

With the results of the sine-sweep analyses the MS (margins of safety) of different structural elements like composite plates, threaded connections, etc. are calculated.

5 RANDOM AND ACOUSTIC EXCITATIONS

While the spacecraft is being transported by the launching vehicle towards the point of separation there are two important actions exerted by the launching vehicle on the satellite:

- random excitations introduced to the spacecraft through its base;
- acoustic excitations.

The random vibrations at the interface launcher-spacecraft are characterized using a PSD (Power Spectral density) distribution; in Figure 19 we present the one corresponding to the SOYUZ launching vehicle.
Figure 19. Random base vibrations *PSD_Launcher* (for qualification of components and for acceptance of components)

The methodology for verifying the different components is:

- define, either experimentally or with finite element models, the transfer function between the base and any component as a function of the frequency: \( TF_{(i-j)}(f) \), where \( i \) is the base vibration direction and \( j \) is the direction of the component vibration;
- calculate for \( j = 1,2,3 \) and for each frequency in the range [20Hz – 2,000Hz]
  \[
  PSD_{\text{Component}}_{jj} = PSD_{\text{Launcher}} \left[ \max(TF_{11-jj}, TF_{22-jj}, TF_{33-jj}) \right]^2
  \]

- knowing the natural frequencies of the component \( (f_n) \) calculate the VRS (vibration response spectrum) for each natural frequency as shown in Figure 20 [4]. For this purpose use Miles equation in the form,

\[
VRS_{\text{Component}}_{jj}(f_n, \xi) = \frac{[1 + (2\xi \rho_i)^2]}{\left[ (1 - q_i^2)^2 + (2\xi \rho_i)^2 \right] \cdot PSD_{\text{Component}}_{jj}}
\]

In the above equation \( q_i = \frac{f_i}{f_n} \) and \( \xi \) is the relative damping.
Santiago Tempone, Marcos Franceschini and Eduardo N. Dvorkin

Figure 20. Determination of the VRS for each natural frequency of the component. PSD_component is determined with Eqn. (2)

- for each component natural frequency the RMS acceleration \( (grms) \) is calculated [4] as the area underneath the corresponding curve (blue area in the case in Figure 21);

The calculated \( grms \) is checked against the component allowable excitation.

In the ARIANE 5 launching vehicle, which is the one that was used for launching the ARSAT-1, it is specified that for less than 100 Hz the random environment is covered by the specified sine sweep environment and that the defined acoustic specification covers the excitations produced by random vibrations for frequencies above 100 Hz.

In Figure 22 it is shown the pressure level that was used in the DAF (direct acoustic field) test.

Figure 21. \( grms \) calculation [4]

Figure 22. Sound Pressure Level for Ariane 5
In Figure 23 it is shown the arrangement for the PFM DAF test with the speakers surrounding the spacecraft and the control microphones indicated.

For assessing the effect of the acoustic field on the spacecraft components it is used as inputs the readings of the accelerometers located at the components basis (or quite close to them) in the DAF test.

The accelerometer readings are transformed into PSD at the base of each component.

The assessment methodology is based on the VRS methodology, using Eqn. (3) onwards.

![Figure 23. DAF test arrangement at CEATSA (Bariloche, Argentina)](image)

6 THERMAL STRESSES

The calculation of thermal stresses proceeds from the determination of the satellite thermal maps for different positions along its orbit; a standard finite element procedure is used with a thermo-elastic material model, except for the threaded connections.

In Section 2 it was discussed the standard method for modeling bolted connections; however, for the thermo elastic analysis, since its results are fundamental for assessing the alignment preservation of the spacecraft in orbit under thermal cycles, it was considered necessary to introduce in the model:

- the possibility of connections slipping;
- the consideration of the hysteresis that may be introduced by the connections slipping.

In Figure 24 a bolted joint is represented. The nonlinear elastic CBUSH that models the bolted joint (Figure 25) includes and initial very stiff elastic part and then the bolted joint slip when the shear force $Q$ is greater than the allowable shear force that can be transmitted by friction between the jointed surfaces. If the shear load is increased after this point, the bolt slides through the clearance. When the bolt gets in touch with the plates a stiffer part of the characteristic Load/Displacement curve represents the bearing between the bolt and the plates.
Slippage occurs when the shear force exceeds the maximum allowable shear force, that is, when

\[ Q > \mu P_0 \]  

(3)

where,
\( \mu \): is the coefficient of static friction between the plates;
\( P_0 \): is the axial bolt preload.

The resulting nonlinear elastic CBUSH curve is displayed in Figure 25.

The limitation of this connection model is that it cannot describe the hysteresis of the connections; hence an elastoplastic model was developed using elastic CBUHES and elastoplastic rod elements as shown in Figure 26 and Figure 27.
Figure 26. Evolution from the nonlinear elastic connection model to the elasto-plastic connection model.

Figure 27. Details of the elasto-plastic connection model.

The Load / Displacement characteristic of this model is represented in Figure 28.
Figure 28. Elasto-plastic connection model behavior

Using this model it was shown that the hysteresis of the bolted connections is negligible and the more simple nonlinear elastic model can be reliably used.

7 CONCLUSIONS
A procedure for the development of finite element models of a spacecraft was discussed. The procedure was used by SIM&TEC and INVAP for the modeling of the ARSAT-1 communications satellite.

Acknowledgement
The authorization of ARSAT S.A. INVAP S.E. for the publication of this paper is gratefully acknowledged.

REFERENCES


COMPUTATIONAL MODELLING: A TOOL TO ADD ECONOMIC VALUE TO THE INDUSTRIAL PRODUCTION

RITA G. TOSCANO\textsuperscript{1}, MARCELA B. GOLDSCHMIT\textsuperscript{1}, SANTIAGO TEMPONE\textsuperscript{2} and EDUARDO N. DVORKIN\textsuperscript{1}

\textsuperscript{1}SIM&TEC S.A.  
Av. Pueyrredón 2130 5\textsuperscript{a}A Buenos Aires, Argentina  
rtoscano@simytec.com  
mgoldschmit@simytec.com  
edvorkin@simytec.com

\textsuperscript{2} INVAP S.E.  
Av. Comandante Luis Piedrabuena 4950, San Carlos de Bariloche, Argentina  
stempone@invap.com.ar

Key words: new technologies, modeling tools, finite elements.

Abstract. The scientific-technological development is essential for the sustained growth of our regions; the production of quality goods with high added value is an important step forward when compared with the production of raw materials. Computational Mechanics is an essential tool for the development of new technologies and for the optimization of the existing ones [1].

The industry faces technological problems increasingly more complex, and the numerical simulation of those technological problems induces scientists to computational developments of greater complexity.

Since technological decisions are reached based on the results provided by numerical models, it is evident that these models have to be highly reliable. Therefore, it is essential that sophisticated modeling techniques are used; that highly qualified engineers develop models and that the results are validated experimentally using industrial or laboratory determinations [2].

The examples used to illustrate this article are taken from real applications developed for industry: the structural verification of the communications satellite ARSAT-1 and the modeling of rock fracturing processes.

1 INTRODUCTION

The scientific-technological development is essential for the sustained growth of our regions; the production of quality goods with high added value is an important step forward when compared with the production of raw materials. Computational Mechanics is an essential tool for the development of new technologies and for the optimization of the existing ones [1].

The industry faces technological problems increasingly more complex, and the numerical simulation of those technological problems induces scientists to computational developments of greater complexity.
Since technological decisions are reached based on the results provided by numerical models, it is evident that these models have to be highly reliable. Therefore, it is essential that sophisticated modeling techniques are used, that highly qualified engineers develop models and that the results are validated experimentally using industrial or laboratory determinations [2].

The modeling process must follow the following steps:

- Identification of the physical phenomenon to be analyzed.
- Formulation of the mathematical model: determine the system of differential equations that best represents the physical phenomenon, defining the appropriate domain, border conditions, initial conditions, etc. At this stage it is necessary to decide what physical aspects are indispensable to take into account in the model and assumptions about the response of the material, loads, friction, etc.
- Development of the numerical model: in most of the cases it is necessary to solve the system of equations in an approximate way, using numerical methods, as for example the method of finite elements (method in which we will focus on this work) [3].
- Verification: in the verification process it must be proved that the equations are resolved correctly, and therefore this is a mathematical step [2]. In this step it must be proved that the numerical scheme is convergent and stable. It is important to notice that the verification process is not related only with a numerical procedure but also with its effective implementation in software, either commercial or developed in-house.
- Validation: The process of validation must demonstrate that the right equations are solved, and therefore is an engineering step [2]. It is validated neither a formulation nor software, but the use of verified software when used by a design analyst in the simulation of a given process. It is necessary to validate the complete procedure.

2 FROM THE PHYSICAL TO THE MATHEMATICAL MODEL

The analyst must understand the physical phenomenon that must model and have enough information on the subject, to include the most relevant features in the model. The expertise of the analyst and a clear definition of expected results are fundamental to the definition of an appropriate mathematical model.

Due to geometric or material nonlinearities most of the models describing physical phenomena of technological importance are non-linear.

In the analysis of a solid under mechanical and thermal loads, some of the nonlinearities found in the formulation of the mathematical model are [4]:

- Geometric nonlinearities
  The equations of equilibrium must be fulfilled in the unknown solid deformed configuration and not in the unloaded known configuration. If these settings are very similar, this nonlinearity can be omitted. An intermediate step would be to consider the balance in the deformed configuration but assuming that deformations are very small. This is also an important simplification in the mathematical model.

- Contact conditions
Contact conditions are unilateral restrictions, contact forces are distributed in an area that is initially unknown to the analyst.

- Material non-linearities
  Elasto-plastic material models (metals), phenomenon of creep with high temperatures, non-linear elastic materials (polymers), fragile materials, changes in phase in solid state, etc.

In the case of fluids or heat transfer, some of the involved nonlinearities are:

- Non-constant viscosity or compressibility: rheological materials and turbulent flows.
- Convective terms of acceleration.
- Temperature-dependent thermal properties: for example changes of phase.

3 EXAMPLES OF TECHNOLOGICAL APPLICATION

The examples used to illustrate this article are taken from real applications developed for the industry.

3.1 Structural models and mechanical tests in the development of a satellite

Computational models were developed to verify the resistance and orbital alignment of satellite Arsat-1, designed and built by INVAP S.E.

The results of computational models were used by the designers of the satellite to predict its dynamic behavior during environmental tests, check its structural integrity and its ability to withstand the thermal cycles in orbit.

Dynamic and thermal vacuum (TVAC) load tests were performed and the results of the simulations were compared with the experimental ones, in order to validate the computational-experimental qualification process of the satellite.

The objective of the dynamic tests was to check the structural integrity of the satellite during launch, while the objective of the thermal test was to demonstrate that the satellite’s equipment are able to withstand the extreme operating temperatures, as well as to validate the thermal mathematical models used to predict in-orbit thermal maps. Finite element models where used to verify the structural integrity of the spacecraft during TVAC tests, and during its entire lifetime in orbit.

Figure 1 shows the geometry and the corresponding finite element model of ARSAT-1 protoflight model (PFM).
Figure 1. Satellite ARSAT-1 PFM, geometry and finite element model.

Figure 2 presents the dynamic test model and the obtained bending second mode. Comparing with the experimental test it was found an excellent correlation between both results (in the order of 5% for structure fundamental modes).

Figure 2. ARSAT-1 PFM. Test configuration and prediction of the second bending mode

In figure 3 it can be observed the temperature data in each sector of the satellite and the respective thermal model data, used to perform the termo-elastic analyses.
3.2 Simulation of hydraulic fracture

The objective of this work is to develop the ability to model hydraulic fracture processes. As the first case of validation, the Brazilian test (diametric compression of a cylindrical sample) was simulated, with a central notch [6, 7].

Different failure criteria were implemented, and the results are shown in figure 4, which compares cracking trajectories obtained experimentally with the numeric ones. The Maximum Tension criterion gave the best results, as it is shown in the figure. The ratio between the numerical failure load and the experimental one is 0.97; therefore this stage was validated.

Figure 3. Termo-elastic analysis of ARSAT-1. From temperature data to finite element model

Figure 4. Modeling of the process of a rock formation fracture
The next stage was to model a rocky massif, loaded with the weight of the rock which is above (figure 5), and to model the hydraulic fracture; figure 6 shows the progress of the fracture.

\[
\sigma_{zz} = \rho g h
\]

Figure 5. Vertical stresses

\[
\begin{align*}
\sigma_{theoretical} &= -0.24525 \text{ MPa} \\
\sigma_{Numerical} &= -0.248 \text{ MPa}
\end{align*}
\]

Figure 6. Pattern of fracture hydraulic pressure (horizontal and vertical plane)

12 CONCLUSIONS

The presented results show that the numerical simulation is a powerful tool in the industry to "add innovative value" to products and processes. The reliability of the models is of utmost importance, so it requires a strongly trained professional to understand the phenomenon, design the model, and judge, with the results, the validity of the assumptions and used models. The experimental validation of the model is indispensable to have a really reliable tool to make technology decisions.

REFERENCES


NUMERICAL SIMULATION APPLIED TO THE TIRE INDUSTRY

JORGE KUSTER$^{1,2}$, BRUNO A. TOSCANO$^{1,3}$, GABRIEL N. CURTOSI$^{1,4}$, Y TOMÁS ARECHAGA$^{1,5}$

$^1$Research Group of FATE Tires Argentina
Blanco Encalada 3003 (B1644GPK) Victoria - Buenos Aires Province
http://www.fate.com.ar

$^2$jkuster@fate.com.ar, $^3$btoscano@fate.com.ar, $^4$gcurtosi@fate.com.ar, $^5$tarechaga@fate.com.ar

Key words: Tire, Finite Element Method, Structural Analysis, Rubber Compound Extrusion

Abstract. The modeling using the finite element method is a powerful tool to analyze a complex structure such as the one of the tires. The models used to evaluate the properties and performance of the tires take into account a high deformation and displacement of the structure. Usually, the materials are considered nonlinear, anisotropic and viscoelastic.

In this work, an overview of some projects developed by the FATE Tires Research Group is shown which involve not only the finished product but also stages of production such as extrusion of rubber compound profiles.

Some of the following applications are addressed: study of the influence of the tread pattern in the tire performance, characteristics functions of a tire obtained by simulation of a tire testing machine, swelling prediction of extruded rubber compounds applied to die design, design of tire mold cavity profiles, simulation of tire tests and modal analysis among others.

Many of the performed simulations were validated with experimental testing.

1 INTRODUCTION

The modeling using the finite element method (FEM) is a powerful tool to analyze a complex structure such as the one of the tires. The models used to evaluate the properties and performance of the tires take into account a high deformation and displacement of the structure. Usually, the materials involved in the tire structure are considered nonlinear, anisotropic and viscoelastic.

In this work, an overview of some projects developed by the FATE Tires Research Group is shown, which involve not only the finished product but also the stages of production such as extrusion of rubber compound profiles.
For reference, the adopted wheel axis system is the TYDEX-C [10], which is shown in Figure 1.

**Figure 1**: Wheel Axis Systems

## 2 CHARACTERISTICS FUNCTIONS OF TIRES OBTAINED FROM FEM

### 2.1 Characteristics Functions

The characteristics functions (also known as "Characteristic shape factors") represent the tire in a simple wheel vehicle mathematical model which describes, in an objective way, maneuverability aspects. Three characteristics functions are defined:

- Normalized Lateral Force (NLF).
- Load Transfer Sensibility (LTS).
- Load Sensibility (LS).

where:

\[
NLF(\alpha, \bar{F}_z) = \frac{F_y(\alpha, \bar{F}_z)}{\bar{F}_z} 
\]

\[
LTS(\alpha, \bar{F}_z) = -\frac{1}{2} \frac{\partial^2 F_y(\alpha, \bar{F}_z)}{\partial \bar{F}_z^2} \bigg|_{\bar{F}_z} 
\]

\[
LS(\alpha, \bar{F}_z) = \frac{\partial F_y(\alpha, \bar{F}_z)}{\partial \bar{F}_z} 
\]
2.2 Finite Element Model

In this case, a 175/65R14 tire was modeled. The model was made using in-house software and the resolution of the problem using the ABAQUS [5] solver. Figure 2 shows the employed mesh.

![Mesh Details of the Model](image)

All the reinforcements of the tire were modeled through the REBAR layer technique available in ABAQUS [5]. The nodes in contact with the rim (highlighted in red in Figure 2) were linked with center nodes using a rigid body constraint, rim constraints were applied to those central nodes. The main aim of this work is to know forces and moments generated by the tire during service, these service conditions means different vertical loads and slip angles (where slip angle is the angle between the rim plane orientation and the wheel forward direction, $\alpha$ in Figure 1). The loads and boundary conditions applied are shown in Figure 3, where it can be seen that for each load condition a slip angle sweep was made.

2.3 Results

Figure 4 shows the lateral force and Figure 5 the moment generated by the tire with the boundary conditions imposed. The obtained curves have the same shape that the obtained with a testing wheel machine, showing a good agreement with the performed numerical simulation.

Figure 6 shows the relationship between the lateral force ($F_y$) obtained from the tire in function of the radial force ($F_z$). It can be seen that the relationship is not linear and this phenomena is the origin of the Load Transfer Sensibility (LTS) concept, this means that in a cornering maneuver the total lateral force obtained in the axis varies depending
on the cornering intensity, being less the total axis force available when the maneuver becomes harder.

With the obtained results it is possible to calculate the characteristics functions. Based on Figure 6 the fitting of the curves was made with a third grade polynomial through the method of least squares, where:

\[ F_y(\alpha_i, F_z) = b(\alpha_i)F_z + c(\alpha_i)F_z^2 + d(\alpha_i)F_z^3 \]  \hspace{1cm} (4)

Operating with (4) in (1), (2) and (3), the next expressions are obtained:

\[ NLF(\alpha_i, F_z) = a(\alpha_i) + b(\alpha_i)F_z + c(\alpha_i)F_z^2 \]  \hspace{1cm} (5)

\[ LTS(\alpha_i, F_z) = -b(\alpha_i)F_z - 3c(\alpha_i)F_z^2 \]  \hspace{1cm} (6)

\[ LS(\alpha_i, F_z) = a(\alpha_i) + 1.8b(\alpha_i)F_z + 2.45c(\alpha_i)F_z^2 \]  \hspace{1cm} (7)

Once determined \( a(\alpha_i) \), \( b(\alpha_i) \) and \( c(\alpha_i) \) the characteristics functions of the tire could be obtained through numerical simulation. In Figure 7 a summary of results is shown.

All this information is useful to determine the performance of different design proposals in the development stage of the tire.

Figure 3: Vertical Load (a) and Slip Angle (b) Applied to the Model
Figure 4: Lateral Force Generated

Figure 5: Auto Aligning Moment Generated

Figure 6: Lateral Force in Function of Radial Force
3 INFLUENCE OF TREAD PATTERNS ON TIRE PERFORMANCE

3.1 Analysis Description

In Section 2 the characteristics functions were obtained and with those results the tire performance could be defined. In this kind of analysis the impact of the tread pattern in tire’s performance is evaluated.

3.2 Finite Element Model

The full 3D model of the tire was made using the Symmetric Model Generation (SMG) technique of ABAQUS [5]. Figure 8 shows the partial 3D mesh employed to obtain the full 3D mesh of the model through SMG.

3.3 Tread Patch Validation

The pressure distribution obtained for the tread patch was validated with experimental results. Figure 9 shows on the left the experimental results and on the right the numerical ones.

Figure 7: Summary of Results for Two Proposed Tires
3.4 Results

As seen on Figure 10, the tread pattern affects the performance of the tire in comparison with the slick tire. Table 1 resumes the obtained results for the Cornering Stiffness Coefficient (CSC), Auto Alignment Coefficient (AAC), Lateral Stiffness ($K_y$) and Torsional Stiffness ($K_{M_z}$), where:

\[
CSC(F_z) = \frac{1}{F_z} \frac{\partial F_y}{\partial \alpha} \bigg|_{\alpha=0}
\]

\[
AAC(F_z) = \frac{1}{F_z} \frac{\partial M_z}{\partial \alpha} \bigg|_{\alpha=0}
\]

\[
K_y(F_z) = \frac{\partial F_y}{\partial y} \bigg|_{y=0}
\]

\[
K_{M_z}(F_z) = \frac{\partial M_z}{\partial \theta} \bigg|_{\theta=0}
\]

<table>
<thead>
<tr>
<th></th>
<th>CSC</th>
<th>AAC</th>
<th>$K_y$</th>
<th>$K_{M_z}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slick</td>
<td>0.467</td>
<td>9.530</td>
<td>8.8</td>
<td>4447.0</td>
</tr>
<tr>
<td>Tread Pattern</td>
<td>0.385</td>
<td>9.063</td>
<td>8.7</td>
<td>4160.5</td>
</tr>
<tr>
<td>Difference [%]</td>
<td>-17.6</td>
<td>-4.9</td>
<td>-0.4</td>
<td>-6.4</td>
</tr>
</tbody>
</table>

Table 1: Results Summary
4 EXTRUSION PROCESS SIMULATION

4.1 Process Overview

During the extrusion process of rubber compounds, the material is forced to flow through a die which has the profile of the desired geometry. The complexity of the process is focused on the swelling that the material undergoes once it leaves the die, being the final geometry different from the one in the die. This swelling is function of the material temperature, viscosity, flow rate, etc. The aim of this kind of simulations is to know the shape of the die that guarantees the desired extruded rubber profile.

4.2 Model and Constitutive Equations

The basic equations to solve are the momentum (12) and mass conservation (13).
\[-\nabla p + \nabla T + f = \rho \dot{a} \quad (12)\]

\[\nabla v = 0 \quad (13)\]

where \( p \) is pressure, \( T \) the extra stress tensor, \( f \) the volumetric forces, \( \rho \) the density, \( \dot{a} \) the acceleration and \( v \) the velocity. For a generalized Newtonian fluid [1] \( T \) results:

\[T = 2\eta(\dot{\gamma})D \quad (14)\]

where \( D \) is the strain rate tensor and the shear rate is defined as:

\[\dot{\gamma} = \sqrt{2\text{tr}(D^2)} \quad (15)\]

the shear viscosity is defined through the Bird-Carreau equation [1]:

\[\eta(\dot{\gamma}) = \eta_\infty + (\eta_0 - \eta_\infty) \left[1 + (\tau \dot{\gamma})^2\right]^{(n-1)/2} \quad (16)\]

The viscoelastic model used was a simplified one applied in the Polyflow solver [1], this model allows the swelling prediction with low computational cost.

### 4.3 Finite Element Model

The domain of the model is subdivided into two subdomains, one corresponds with the die and the other with the free surface at the die outlet, here is where the rubber compound swells. Figure 11 shows the two domains and the boundary conditions considered.

![Figure 11: FEM Model Domain and Boundary Conditions](image-url)
4.4 Results

Figures 12 and 13 show the velocity field obtained at the inlet and outlet of the die. Figure 14 shows the pressure distribution at the die wall. The interface between the three materials (cap, base and wing) can be obtained too, as seen in Figure 15. Finally, the numerical result is compared with the real profile obtained in the extrusion line, using a laser profile scanner, showing a very good agreement (Figure 16).

Figure 12: Velocity Field - Top: Die Inlet / Bottom: Die Outlet

Figure 13: Velocity Field at Die Outlet

Figure 14: Pressure Distribution at Die Wall
5 CONCLUSIONS

FEM models are a very useful tool to develop and improve productive process and product performance in the modern tire industry. A huge quantity of simulations could be performed to evaluate the designed tires before its manufacturing, allowing a considerable cost reduction on new projects. Through these techniques, problems of tire durability and manufacturing defects could be covered too. The increase of computational power allows the development of automatic processes where, using mathematical optimization algorithms combined with FEM simulations, a tire performance target is found through design variables, e.g., tire’s profile shape. As a quick overview, these applications were performed using numerical tools to develop and improve tires, among others:

- Tire Performance:
  - Tire forces and moments generation (Drum Test).
  - Static and dynamic contact patch.
  - Development of tire profiles.

- Environmental Issues:
  - Rolling resistance calculation (fuel consumption reduction).
  - Acoustics analysis (external and internal noise reduction).

- Vehicle Comfort:
  - Non-uniformities evaluation.
  - Dynamic impact.
  - Modal analysis.

- Tire Security and Durability:
- Braking distance improvement.
- Hydroplaning.
- Thermal distribution during service.
- Endurance test.
- Plunger energy test.

- Manufacturing Process:
  - Rubber compound extrusion (Swelling).
  - Tire manufacturing process.

REFERENCES


A REVIEW ON THE NONLINEAR DYNAMICS OF PENDULUM SYSTEMS FOR ENERGY HARVESTING FROM OCEAN WAVES

FRANCO E. DOTTI\textsuperscript{1}, FLORENCIA REGUERA\textsuperscript{2} AND SEBASTIÁN P. MACHADO\textsuperscript{3}

\textsuperscript{1}Grupo de Investigación en Multifísica Aplicada (GIMAP), UTN-FRBB. Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). 11 de Abril 461, B8000LMI Bahía Blanca, Argentina. fdotti@frbb.utn.edu.ar

\textsuperscript{2}Centro de Investigaciones en Mecánica Teórica y Aplicada (CIMTA), UTN-FRBB. Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). Departamento de Ingeniería, UNS. 11 de Abril 461, B8000LMI Bahía Blanca, Argentina. freguera@frbb.utn.edu.ar

\textsuperscript{3}Grupo de Investigación en Multifísica Aplicada (GIMAP), UTN-FRBB. Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). 11 de Abril 461, B8000LMI Bahía Blanca, Argentina. smachado@frbb.utn.edu.ar

Key words: Parametric Pendulum, Wave Energy, Nonlinear Dynamics.

Abstract. In this article we explore the state of knowledge in the field of nonlinear rotatory dynamics of pendulum systems, with a view in energy harvesting from ocean waves. Although the idea of extracting energy from the ocean by pendulum systems is recent, there exist a previous wealth of analytical, numerical and experimental research which was aimed to understand the dynamics of the parametric pendulum. The key points of this research are reviewed, and also the still many unanswered questions are discussed.

1 INTRODUCTION

The world’s wave energy potential is estimated to be around 2.5–3 TW [1]. Recent evaluations of the Atlantic coast of South America estimate available wave energy within the order of 200 GW, while for the Pacific coast the assessment arises to 320 GW. Wave energy is currently being used in developed countries: about 55 TWh per year of energy consumption are replaced by energy coming from wave motion in the United States, while in Europe this value corresponds to 280 TWh. Given these trends, it is logical to think that the use of wave energy presents a huge viability in the foreseeable future.

Systems which convert wave energy into rotational kinetic energy of a parametric pendulum seem to have a great potentiality. The idea consist of a pendulum on a floating platform. The waves induce vertical motion to the system at a given frequency. If one success to maintain stable rotations, a generator attached to the axis of rotation of the pendulum may be able to extract electrical energy. While being conceptually simple, the development of these systems is at an embryonic stage [2–4], and there is still no full-scale prototype available. The main reason for this lies in the practical difficulty of achieving and maintaining rotational motion.
under the irregular excitation of waves. Therefore, active control is needed to set the desired period 1 rotation irrespective of the initial conditions, and to stabilize this motion for changing forcing parameters.

The nonlinear dynamics of the pendulum is one of the most studied topics of physical science of all time. It is not possible to review all contributions and hence we only focus on rotational motion of the vertical parametric pendulum, which is the one associated with energy harvesting. This article is organized as follows. The next section gives an overview of the dynamical behavior of the vertical parametric pendulum. In sections 3 and 4 the most relevant contributions to the understanding of the dynamics of rotating motions are reviewed. Finally, in section 5, energy harvesting from ocean waves using pendulum systems is explored. The final summary includes a global analysis and a discussion about the forthcoming challenges.

2 THE VERTICAL PARAMETRIC PENDULUM

The vertical parametric pendulum (Fig. 1) is a classic example of a universal problem in physical sciences, which is the escape from a potential well [5]. The equation of motion of the vertical parametric pendulum can be set up by using Lagrange’s equation for single-DOF non-conservative systems, and its derivation can be easily followed in any classic book of nonlinear dynamics [6,7]. It is a second-order ordinary differential equation, which is given by

\[ ml^2 \frac{d^2 \theta}{d\tau^2} + c \frac{d\theta}{d\tau} + ml \left( \frac{d^2 y}{d\tau^2} + g \right) \sin \theta = 0 \] (1)

where \( m \) is the mass of the pendulum bob, \( l \) the distance between the center of gravity and the axis of the pendulum, \( c \) the viscous damping coefficient, \( \tau \) the time, \( g \) the acceleration of gravity, \( y(\tau) \) the vertical displacement of the axis, and \( \theta \) the angle measured from the hanging position.

Figure 1: The parametrical pendulum, excited in vertical direction.

The natural frequency of the system for small oscillations is \( \omega_0 = \sqrt{g/l} \). If the excitation is assumed to be harmonic, i.e. \( y(\tau) = -y_0 \cos(\Omega \tau) \), and after dividing by \( ml^2 \), (1) becomes
By defining

\[ t = \omega_0 \tau, \quad \beta = \frac{c}{\omega_0 ml^2}, \quad \omega = \frac{\Omega}{\omega_0}, \quad p = \frac{y_0 \Omega^2}{l} \]  

(3)

and substituting in (2), we obtain the rescaled equation of motion

\[ \ddot{\theta} + \beta \dot{\theta} + \left[ 1 + p \cos(\omega \tau) \right] \sin \theta = 0 \]  

(4)

where the dot means the derivative with respect to dimensionless time \( t \). In (4), the magnitudes \( \beta, p \) and \( \omega \) are the parameters associated respectively to damping, forcing amplitude and forcing frequency. Depending on the settings of these parameters and on the choice of initial conditions \( \theta_0 \) and \( \dot{\theta}_0 \), the steady state of (4) may correspond to different kind of motions, namely:

- The rest or equilibrium position, \( (\theta, \dot{\theta}) = (2k\pi, 0), k \in \mathbb{N} \): the pendulum bob simply moves in the same direction of the load (i.e. vertical direction).
- Oscillating motion: the angle do not exceed \( \theta = \pm \pi \).
- Rotating motion: the bob make revolutions about the pivot point.
- Chaotic motion (Fig. 2): the trajectories rotate clockwise and anticlockwise in chaotic fashion. This motion is often called tumbling chaos [8].

\[ \text{Figure 2: Tumbling chaos of the vertical parametric pendulum. } p = 1.1, \omega = 1.56, \beta = 0.1, \left( \theta_0, \dot{\theta}_0 \right) = (\pi/3, -3): \]

(a) phase portrait (projection of the tridimensional phase space) showing a tangled combination of oscillations and rotations; (b) the associated Poincaré map is a strange attractor, confirming chaotic motion [5].

The rest position is always a solution of Eq. (4), either stable or unstable. The graphic of Fig. 3 has been constructed numerically by evaluating the density of occurrence of the rest as a steady state stable solution, for several initial conditions. The resonance zones are those regions of the parameter space where the rest position is unstable, and they occur at \( \omega = 2/m, m \in \mathbb{Z} \). These zones have been widely studied since they indicate which combinations of parameters have the rest position as a steady state solution [5]. In Fig. 3, the damping parameter \( \beta \) has been
fixed: it has been demonstrated that increasing damping moves the resonance zones upwards [9-12]. For $\beta = 0$, the vertices of the resonance zones coincide with the horizontal axis ($p = 0$).

**Figure 3:** Stability diagram showing the resonance zones (represented as blue tongues) where the rest position is unstable. The graphic was constructed numerically as a density contour of the occurrence of the rest solution ($\beta = 0.1$). Blue denotes zero, indicating that the steady state of the system cannot be the rest position.

**Figure 4:** Phase portraits of the steady state orbits which include rotating motion, as classified in [9]: (a) pure rotation for $p = 0.9$, $\omega = 2.7$; (b) oscillating rotation for $p = 1.5$, $\omega = 0.19$; (c) straddling rotation for $p = 1.4$, $\omega = 0.15$; (d) large amplitude rotation for $p = 1.7$, $\omega = 1.0$. In all cases $\beta = 0.1$ and $(\theta, \dot{\theta}) = (\pi/3, 0)$.

The steady state orbits including rotating motion were classified by Garira and Bishop [9] in four categories: pure rotations, oscillating rotations, straddling rotations and large amplitude
rotations (Fig 4). Among all of these solutions, \textit{pure rotations} have an important attribute: the angular velocity maintains the same sign for all time, either $\dot{\theta}(t) > 0 \ \forall t$ or $\dot{\theta}(t) < 0 \ \forall t$. This ensures that there is no change of direction in the course of rotation, i.e. there is no oscillating motion (in counterpart, for pure oscillations we have $-\pi < \theta(t) < \pi \ \forall t$). Pure rotations always exist in conjugate pairs: clockwise and anticlockwise. For energy harvesting purposes, a pure rotation is the desired motion and it is regarded as a synonym of “rotation”, while the other categories presented in [9] are considered merely as oscillations.

For a specific set of parameters, the coexistence of periodic and chaotic solutions is possible. The parameter space of Fig. 5 shows the combination of motions. For low excitation amplitudes, the rest position is the commonest solution; by increasing $p$, oscillations, rotations (denoted as “Rot$_{1}^{1}$” and “Rot$_{2}^{2}$”) and chaos appear. It is worthwhile to note that rotations are only possible for approximately $\omega > 1$. It can be seen that pure rotations are the dominant type of stable rotating solutions in the main resonance zone ($\omega = 2$). The resonance zones presented in Fig. 3 can also be observed in Fig 5.

\textbf{Figure 5:} Parameter space showing, approximately, all the possible combinations of solutions ($\beta = 0.1$). The map may have some minor inaccuracies due to the choice of initial conditions and/or inability of the solution to escape of the transient. Rot$_{2}^{2}$ means “pure rotations of period 2 or higher”.

If the response has to be studied by increasing the forcing magnitude $p$ within the main resonance zone ($\omega = 2$) we can consider the numerical bifurcation diagram of Fig. 6. Initially the only stable solution is the rest position. For $p \approx 0.2$ we enter inside the tongue, and the rest position loses stability by a Hopf bifurcation (supercritical), creating a period 2 oscillation [5]. Up to $p \approx 0.45$, the oscillatory attractor is the only solution: all the trajectories stay in the potential well. Beyond that value, a pair of symmetric rotations appear by a fold (saddle-node) bifurcation: some trajectories are now allowed to escape from the well. At $p \approx 0.88$, a pair of period 3 rotations born also by folds, but soon they run into a period doubling cascade (supercritical flip bifurcations) until vanish in a crisis at $p \approx 0.98$. Period 1 rotations also double the period at $p \approx 1.35$ and run into cascade at $p \approx 1.72$. For $p$ increasing further, the symmetric attractors lose stability in respective crises at $p \approx 1.81$. Beyond this threshold, tumbling chaos and oscillating solutions take place. Windows of periodicity are possible into chaos, as the one at $p \approx 2.6$ corresponding to a period 8 large amplitude rotation [9]. Useful behavior charts of
these bifurcation mechanisms were presented by Szemplinska-Stupnicka et al. [13].

Figure 6: Bifurcation diagram of the angular velocity for $\omega = 2$, with $\beta = 0.1$. (─): pure clockwise rotations, (─): pure anticlockwise rotations, (─): rest, (─): oscillations and oscillating-rotating motions, (─): tumbling chaos.

The dynamics of the parametric pendulum can be very complex. The subject of this article is the occurrence of rotations, which represent the most important motion with a view in energy harvesting. This study starts in the following section.

3 THE SEARCH FOR ROTATING PERIODIC ORBITS

Aside from some early discussions and experiments [14,15], and some tangential references about chaotic motions [16,17], the first researchers to study rotary solutions in depth were Leven and co-workers [11]. Varying damping and amplitude of excitation, they experimentally determined threshold values for period doublings of the main rotational motion, and for chaos. They show that the curves in the $p$-$\beta$ plane were monotonously increasing with $\beta$, i.e. that increasing damping moves the resonance zones upwards in the $p$-$A$ plane (see Fig. 3). They also noted that the coexistence of period 1 and 3 rotating attractors was possible.

Fully focused on rotating motion, Clifford and Bishop [8] made a categorization of rotations according to their period $n$ (i.e. the number of load cycles needed to obtain one cycle of the response), and the number of complete rotations performed in $n$ periods of forcing, $r$. They used the notation $(n, r)$, and presented a numerical study of the bifurcational behavior for the purely rotating (1,1) orbit, displaying their results in the parameter space $p$-$\omega$ of Fig. 7. They accurately identified some important events: the onset of (1,1) rotation (Line A), the threshold of period-doubling to (2,2) rotation (Line B), the end of period-doubling cascade in a crisis (Line F) and a reverse period-doubling cascade where the (1,1) solution re-establishes after tumbling chaos (Line U). One may compare Fig. 7 with Fig. 5, which was prepared for the present article and took several days, and remark the precision achieved in [8] with a computer twenty years older.

Clifford and Bishop [8] showed how to find pure rotations (Fig. 7). But as we have seen previously (Fig. 5-6), rotations can coexist with other motions. Hence, the following question can be made: how difficult is it to obtain a steady state of pure rotation with a random initial condition? In this direction, Xu et al. [18] studied numerically some bifurcation scenarios for varying $p$ within the main resonance zones $\omega = 2$, $\omega = 1$ and $\omega = 2/3$. For each value of the
varied parameter the same initial conditions were used, thus allowing a comparison. They concluded that “pure rotations can be easily obtained in the main resonance zone, but can rarely be found in the higher order resonance zones”. This qualitative conclusion reveals an important fact: rotations are not a rare phenomenon in a parametric pendulum.

![Parameter space for the clockwise purely rotating orbits (β = 0.1) presented in [8].](image)

**Figure 7:** Parameter space for the clockwise purely rotating orbits (β = 0.1) presented in [8].

![Basin erosion sequence for ω = 2 and β = 0.1 as p is increased; (a) p = 0.420, a pair of symmetrical period 1 rotating attractors born by a fold; (b) p = 0.655, the basin of rotations grow into the basin of oscillations; (c) p = 0.655, a homoclinic tangency occur and fractal erosion of the oscillatory basin has started; (d) p = 0.900, period 3 rotating attracts born by a fold and contribute to the erosion of the oscillatory basin; (e) p = 1.331, pure oscillatory is almost fully eroded; (f) p = 1.750, the rotating basins are also almost fully eroded, tumbling chaos is imminent. (─): clockwise rotations, (—): anticlockwise rotations, (—): oscillations, (∼): tumbling chaos.](image)

**Figure 8:** Basin erosion sequence for ω = 2 and β = 0.1 as p is increased; (a) p = 0.420, a pair of symmetrical period 1 rotating attractors born by a fold; (b) p = 0.655, the basin of rotations grow into the basin of oscillations; (c) p = 0.655, a homoclinic tangency occur and fractal erosion of the oscillatory basin has started; (d) p = 0.900, period 3 rotating attractors born by a fold and contribute to the erosion of the oscillatory basin; (e) p = 1.331, pure oscillatory is almost fully eroded; (f) p = 1.750, the rotating basins are also almost fully eroded, tumbling chaos is imminent. (─): clockwise rotations, (—): anticlockwise rotations, (—): oscillations, (∼): tumbling chaos.
In quantitative terms, the answer to the previous question comes from the fractal erosion of the basin of attractions. This is a typical phenomenon of systems involving escape from a potential well. It is triggered by the homoclinic tangency, which is the equivalent event of a homoclinic bifurcation in a Poincaré map [5]. In Fig. 8 the basins of attraction at increasing levels of $p$ have been determined numerically using a grid in the $(\theta_0, \dot{\theta}_0)$ starting space. After the homoclinic tangency ($p = 0.655$), “fractal fingers” are created which sweep dramatically across the basin of the oscillatory attractor as $p$ increases. At a critical value of the forcing magnitude ($p = 1.342$, according to [19]) the oscillations are destroyed. The study of these interactions among attractors can be addressed by an integrity analysis, in which magnitudes associated to each basin of attraction are quantified and compared [20].

This is what Lenci and Rega did [19] for the main resonance case (Fig. 6). Their goal was to highlight how the integrity of the system is modified by varying $p$. They proposed a new integrity indicator: the integrity factor (IF), which is the normalized radius of the largest circle entirely belonging to a basin. Their main results are reproduced in Fig. 9. The competition between the in-well (oscillating) and out-of-well (rotating) attractors is evident: R1 basins grow up against the O2 basin, whose integrity is reduced up to its disappearance. The sharp fall in O2 integrity (labeled as A) starts at $p = 0.655$ with the homoclinic tangency (see Fig. 5c) and corresponds to the fractal erosion of the O2 basin: the IF of O2 basin is strongly reduced by the fractal fingers. Since the total area of the O2 basin is almost unchanged during the erosion, the classical integrity profiles based on area measurement [20] practically do not record this meaningful event. It can be also noted in Fig. 9 that R3 attractor drastically reduce the integrity of R1. This is because R3 basins appear inside R1 basins, thus reducing the value of its IF.

There are many more articles with experimental and numerical work, in which the dynamics of the parametric pendulum is addressed. As a matter of space, we only selected those we have considered more relevant in terms of rotating solutions.

4 FINDING PRACTICAL FORMULAS

It is known that analytical work can be a difficult task in nonlinear dynamics. However, by means of perturbation methods, some researchers obtained explicit formulas which deserve to
be reviewed by virtue of their practicality. Koch and Leven [12] derived an analytical solution for the threshold of homoclinic tangency. This was done by applying the Melnikov method, (a global method within the perturbation theory, see [21] for more detail). They obtained

$$p_{\text{hom}}(\omega, \beta) = 4\beta \pi^{-1} \omega^{-2} \sinh \left( \frac{\pi \omega}{2} \right)$$

Due to its simplicity, Eq. (5) has been widely used to predict the onset of the erosion of the oscillatory basin of attraction [19,22,23]. Moreover, by virtue of (5) the homoclinic tangency is the only global event which can be predicted analytically in an accurate form [5].

Also applying the Melnikov method, Koch and Leven [12] made another prediction: the occurrence of the subharmonic solutions (which can arise by flip or fold bifurcations, as we have seen before). For rotatory solutions they obtained

$$p^{(m)}_{\text{rot}}(\omega, \beta) = 4\pi^{-1} \omega^{-2} \beta k E(1/k) \sinh \left[ \omega k^{-1} K^*(1/k) \right]$$

where \(m = 1, 2, 3, \ldots\), \(K(k)\) denotes the complete elliptic integral of the first kind, \(K^* = K(k')\) with \(k'^2 = 1 - k^2\), \(E(k)\) is the complete elliptic integral of the second kind and \(k\) can be obtained numerically from the resonance condition for the rotating motion

$$2k^{-1} K(1/k) = 2\pi \omega^{-1} \left( m/n \right)$$

being \(m\) and \(n\) relatively prime natural numbers. Eq. (6) was tested against some selected numerical results showing a good agreement for the onset of rotatory motion \((m = 1)\). Only a discussion is present for attractors of higher period but they did not make comparisons.

Lenci and co-workers [22] derived his own version of \(p^{(1)}_{\text{rot}}(\omega, \beta)\), i.e. the curve in the \(p-\omega\) plane of the fold bifurcation, which gives birth to the period 1 rotations. As well as in reference [12], these authors applied a perturbation method and thus their solution is valid for small values of \(p\) and \(\beta\). Considering periodicity and synchronicity conditions, they obtained

$$p^{(1)}_{\text{rot}}(\omega, \beta) = \beta \int_0^\pi f_0(\theta) d\theta \left\{ \int_0^\pi \sin \left( \omega \int_0^\theta \frac{d\theta'}{f_0(\theta')} \right) \sin \theta d\theta \right\}^{-1}$$

The authors considered that for pure rotating orbits \(\dot{\theta}(t) > 0\ \forall t\) is valid, therefore the inverse function \(t = t(\theta)\) of \(\theta(t)\) exists, and then it is possible to define unambiguously the function \(\dot{\theta} = \dot{\theta}(t(\theta)) = f(\theta)\). Thus, in Eq. (8), \(f_0(\theta)\) is the classical solution for the unperturbed pendulum. A further simplification of (8) was performed by means of a Padé approximation and the following formula was obtained

$$p^{(1)}_{\text{rot}}(\omega, \beta) \approx 2\beta \left( 1 + \omega^2 + \omega^4 \right) \left( \omega + \omega^3 \right)^{-1}$$

The authors also gave expressions to calculate the initial conditions \(\dot{\theta}_0\) and \(t_0\), while assuming \(\theta_0 = 0\). An expression for the threshold of the first period-doubling of rotating solution, i.e. \(p^{(2)}_{\text{rot}}(\omega, \beta)\), was also derived in [22] from a stability analysis. This expression is given by

$$p^{(2)}_{\text{rot}}(\omega, \beta) = \left( \frac{1}{\beta} - \frac{1}{\omega} \right) p^{(1)}_{\text{rot}}(\omega, \beta) \left\{ \omega \int_0^\pi \frac{d\theta}{f_0(\theta)} \int_0^\pi f_0(\theta) d\theta \right\}^{-1}$$
While (9) had a good correspondence with numerical results, (10) produced a poor agreement. The authors gave a reasonable explanation in terms of stability: for a flip bifurcation to occur, one eigenvalue has to reach (for growing $\epsilon$) the value $-1$. But this is “far” from the starting point (for $\epsilon = 0$), which is $1$; therefore “large” values of $\epsilon$ would be needed and the perturbation analysis fails. However, it is shown that $P_{rot}^{(2)}(\omega, \beta)$ underestimate the real threshold, so (9) and (10) are a valid conservative estimation of the stability region of the period 1 rotating solutions. Eq. (9) and (10) are very useful in practical terms: depending on the parameters of the system, they indicate a priori if once reached, the rotation will sustain as a steady state motion.

5 ENERGY HARVESTING FROM OCEAN WAVES

While energy extraction is implicit in many of the works discussed so far, it does not represent a central matter. Only in recent investigations, energy harvesting from ocean waves using pendulum systems has been shown as a tangible topic which deserves to be developed.

![Figure 10](image)

**Figure 10**: Map where rotations can be found experimentally according to [23].

![Figure 11](image)

**Figure 11**: According to [24]: (a) Optimum trajectory for different values of damping; ($-$): $\beta = 5 \cdot 10^{-2}$, ($-$): $\beta = 5 \cdot 10^{-4}$, ($-$): $5 \cdot 10^{-2}$. (b) Power input to the system as a function of damping (note that negative power input indicate positive power output from the system). For both figures, $\omega = 1.8$ and $\delta = 0.3$.

In a very interesting article, Lenci and Rega [4] concluded that main stable rotations are possible only in a subset of the theoretical region delimited by (9) and (10) [22]. This is clear
from observing the results of their experiments in Fig. 10. The experimental threshold of rotations, \( p_{\text{rot}}^{(1,\text{exp})} \), can be approximated by joining the square dots, while the proper can be done for the experimental end of stable period 1 rotations, \( p_{\text{rot}}^{(2,\text{exp})} \), by joining the circle dots. In order to explain the results, the authors performed an IF integrity analysis [19] (similar to that of Fig. 9). By considering that rotations can be achieved practically only for high IF values, the authors conclude that the theoretical threshold will never be obtained, even in extremely careful and controlled experiments, since the basins of rotations are not robust enough.

Nandakumar et al. [24] searched the optimal rotational solution which allows maximum energy extraction. The equation considered in this work, analogous to (4), is given by

\[
\theta + \beta \theta - \left[ 1 + \delta \sin(\omega t) \right] \sin \theta = T(t)
\]

(11)

where \( \delta \sin(\omega t) \) is the vertical motion imposed to the pivot, and \( T(t) \) is a control torque added to the system. If \( T(t) > 0 \), the torque represents input energy as an active control, while if \( T(t) < 0 \), it represents energy extracted from the system. The minus sign in (11) is because the authors inverted the direction of gravity. Using (11) and the definition of mechanical power, they proposed an objective function representing the power supplied to the pendulum as

\[
\bar{W} = \frac{1}{2} \pi^{-1} \omega \left\{ \int_0^{2\pi/\omega} \theta \left[ 1 + \delta \sin(\omega t) \right] \sin \theta + \beta \dot{\theta} \right\} dt
\]

(12)

where the solutions \( \theta(t) \), \( \dot{\theta}(t) \) and \( \ddot{\theta}(t) \) for period 1 rotation were assumed as power series with unknown coefficients. With these coefficients as parameters, they minimized (12) to obtain an optimal rotation, which is shown in Fig. 11(a). A surprising result was found for the undamped case: the optimal rotation is a piecewise constant solution regardless of the excitation magnitude and frequency. The authors deduced that if this optimal solution could be achieved, it would be a power output gain of about 27% with respect to a uniform rotating solution.

![Figure 12](image_url)

**Figure 12:** Experimental phase portraits of the response of the pendulum excited by a stochastic wave, obtained by Najdecka and Wiercigroch [3] (a) with velocity control applied; (b) with delayed-feedback control applied.

Of course, the infinite velocity during transitions is unphysical, and it is a consequence of the lack of resistance, allowing the torque to be unlimited. The presence of damping, while smoothening the optimal solution, also dissipates energy. In Fig. 11(b), the objective function
is evaluated for different levels of damping. One can see that if damping is too large, no power can be extracted from the system. If damping decreases, the maximum power which can be extracted approach a limiting value.

In a recent work, Najdecka and Wiercigroch [3] showed experimentally that a stochastic wave can successfully excite the parametric pendulum to rotate. Besides vertical forcing, the authors also regarded horizontal forcing, considering the equation

$$\ddot{\theta} + \beta \dot{\theta} + \dddot{x} \cos \theta + (1 + \ddot{y}) \sin \theta = 0 \quad (13)$$

To represent the time history of excitation, the authors employed the Pierson-Moskowitz formulation [25], which is the most common spectral representation of a wave. Since for the stochastic excitation the pendulum does not follow a single rotational orbit (i.e. “period 1 rotation” cannot be identified), the authors introduced a new variable to identify rotations: the rotational number, given by

$$\lim_{\tau_2 \to \infty} \frac{1}{\omega (\tau_2 - \tau_1)} \int_{\tau_1}^{\tau_2} \text{sgn}(\theta \mod (2\pi)) \dot{\theta} d\tau \quad (14)$$

where $\omega$ is the forcing frequency and $\tau_2 - \tau_1$ is a time span. If $<\dot{\theta}>/\omega \geq 1$ the response is mainly rotational, while if $<\dot{\theta}>/\omega < 1$ the response is oscillatory. This number can be interpreted as a measure of the efficiency of energy exchange between the platform and the pendulum. From the results of experiments and simulations, this researchers found that the rotational number is maximized by introducing horizontal motion to the platform. This conclusion implies an interesting fact: planar excitation can maintain rotations easier than vertical excitation.

We have seen previously that knowing the forcing parameters for a simple harmonic excitation suitable initial condition ensure rotational motion. But with stochastic excitation, active control is crucial to achieve and maintain rotation, since the forcing parameters are variable. Two control techniques were studied in [3] for the parametric pendulum. One is the delayed-feedback method [26], consisting of stabilizing the system by supplying a torque proportional to the difference between the present and a delayed state, which is

$$u = k \left( \theta(t - \tau) - \dot{\theta}(t) + 2\pi \right) \quad (15)$$

where $k$ is a proportionality constant and $\tau$ is the period of the desired periodic response. For the period 1 rotation, $\tau$ can be determined from the period of excitation; but this is hard to identify with a stochastic forcing. This is the reason for the study of the other technique: the velocity control, with a signal independent of the forcing parameters. This control is activated if the instantaneous angular velocity of the pendulum drops below a minimum value $\dot{\theta}_{\text{min}}$, which is minimum velocity necessary for the pendulum to overtake $\pi$ position (i.e. to achieve rotation). Then, the velocity control signal is given by

$$u = k \text{sgn}(\theta) \left( \dot{\theta}_{\text{min}} - \dot{\theta}(t) \right) \quad (16)$$

From numerical studies, the researchers conclude that although both control methods success on maintaining rotation, velocity control would require less energy input to the system. In experimental tests both methods maintain rotations, but velocity fluctuates, resulting in quasiperiodic motion, as shown in Fig. 12. They also report problems with the real time
6 CONCLUSIONS

We attempted to review the key points in the field of nonlinear rotatory dynamics of the vertical parametric pendulum, with a view in the possible application of this system in energy harvesting from ocean waves. Having explored, as perceived by the authors, the main contributions in the scientific literature, it is clear that pendulum systems have a promising future in wave energy extraction. But it is also clear that there is a long way to go until we can see fields of pendulum systems floating over the ocean.

Almost all of the models reviewed consider the parametric pendulum on a rigid base, i.e. a single-degree-of-freedom model. But the structural interaction between the pendulum and its platform has an important effect in terms of stiffness and damping. Therefore, multi-degree-of-freedom systems must be deeply studied. In the case of multi-pendulum systems, the interaction pendulum-base-pendulum is centrally involved in the synchronization mechanism [27]. Synchronization of parametric pendulums under stochastic wave excitation still has not been addressed.

Active control has been proven to be a necessary aid to maintain rotational motion under stochastic excitation [3]. But the investigations are in an early stage, and a suitable optimum controller still needs to be developed in order to ensure the minimum energy input to the system: different control techniques need to be adapted, tested, and compared in order to make energy extraction viable.

In addition, the technology needed for the conversion of the kinetic energy of the pendulum into electrical energy must also be settled and optimized, including devices allowing the storage of the energy extracted. This of course represent a major area of research itself.

Undoubtedly, there are many challenges to face until energy harvesting from ocean waves by parametric pendulums could be a viable, practical, and commonly implemented technology. But under the global requirements of clean energy, we think that these challenges can be soon successfully overcome.

ACKNOWLEDGEMENTS

The authors would like to thank the support of CONICET and Secretaría de Ciencia y Tecnología of Universidad Tecnológica Nacional.

REFERENCES


ANALYSIS OF SELF-BALANCED THERMAL STRESSES IN LONG FLAT RECTANGULAR ISOTROPIC PLATES

TIAGO S. MENDONÇA¹, GABRIEL O. RIBEIRO² AND MARCO A. VECCI³

¹Department of Structural Engineering - Federal University of Minas Gerais
Belo Horizonte - Brazil
tsmrovao@yahoo.com.br

²Department of Structural Engineering - Federal University of Minas Gerais
Belo Horizonte - Brazil
gabriel@dees.ufmg.br

³Department of Structural Engineering - Federal University of Minas Gerais
Belo Horizonte - Brazil
vecci@dees.ufmg.br

Key words: Thermal Structural Analysis, Self-Balanced Thermal Stresses, Thermal Strains.

Abstract. The presence of nonlinear temperature distributions, very common in structures such as spaceships, supersonic aircrafts, industrial furnaces, among others, induces the emergence of self-balanced thermal stresses. However, a unified procedure that allows a rapid and efficient determination of thermal stresses in structures under adverse temperature variations is rarely found in technical literature. This article aims to present the formulation for the determination of self-balanced thermal stresses and thermal strains in long flat rectangular plates that are isotropic and free of external loads at their ends, as well as the comparison of the results of this formulation, with results obtained through the application of the finite element method. The plate in question is extremely thin, thus, it is considered that the plate is working in plane stress state. The expressions for the stress and strain fields are obtained through a very well-known equation in the literature. In this analysis, it will be considered basic simplifying assumptions arising from the condition of distribution of self-balanced thermal stress, where the resulting efforts are zero in the plate cross section and, in addition, it will be considered an approximate temperature field, in polynomial format, which will represent the nonlinear temperature distribution acting along the width of the plate. Thus, the stress and strain fields shall be determined using the polynomial coefficients related to the approximate polynomial field that represents the temperature distribution. The comparison of the approximate analytical results, obtained for the thermal stress and thermal strains, with the corresponding results from the numerical analysis by the finite element method, showed a high level of accuracy for points away from the free ends. Therefore, based on the methodology proposed in this work, the stress and strain fields, arising from a nonlinear thermal variation, can be calculated quickly and accurately using the coefficients of the polynomial that represents the approximate temperature field acting in the plate.
1 INTRODUCTION

During the twenty century there was the advent of several structures that are commonly subjected to nonlinear temperature fields. Structures such as spaceships, supersonic and hypersonic aircrafts, plates which suffer thermal treatments in the industry, among others, are constantly subjected to these fields of temperature. In this way, it has become important the measurement of the stress and strain fields caused by the different thermal gradients found in these structures. However, this area is still little studied and a unified procedure that allows the rapid and efficient determination of stresses and strains in structures under adverse temperature variations is rarely found in technical literature.

Some recent articles are addressed to determine the thermal stresses and thermal strains in plates and beams that are free of external loads at their ends. Articles such as those presented by [1,2] calculate the stresses and strains in a semi-infinite rectangular beam and [3,4] the stresses and strains in thin plates are determined.

In this way, this article aims to present the formulation for the determination of self-balanced thermal stresses and thermal strains in long flat rectangular plates that are isotropic and free of external loads at their ends. These plates have the length substantially greater than the width and are assumed that the plates are requested in the plane stress state.

2 STRESS AND STRAIN FIELDS ACTING IN THE PLATE

In this section, the self-balanced thermal stresses and thermal strains on a long flat rectangular plate that is isotropic and free of external loads at their ends have been determined. For the formulation of the problem it has been considered the flat rectangular plate shown by FIG. 1. It was considered that the plate was requested in the plane stress state by the fact of its thickness, $2b$ dimension, has been considered to be very small with respect to width and the length, $2h$ and $2l$ dimensions, respectively. The plate was subjected to a nonlinear temperature distribution $T=T(y)$ along its width, and it was assumed that the length of the plate was significantly greater than its width.

![Figure 1: Long flat rectangular plate in the plane stress state](image)

The determination of the stress and strain fields acting on the plate of FIG. 1 was performed using the expressions developed by [5]. In [5] the solution developed in [6] was used, obtained for beams, to develop expressions that represent the stress and strain fields acting on the plate described by FIG. 1. In [5] it was assumed a generalized polynomial temperature distribution along the width of the plate and it was found, from the coefficients of the polynomial temperature field, expressions for the stress and strain fields. The same procedure adopted by [5] was performed in [7] for the calculation of the thermal stresses and
thermal deflections in beams.

The solution obtained in [6] for the component of thermal stress $\sigma_x$ acting in a beam is given by:

$$\sigma_x = -E\alpha T + C_1 y + C_2$$  \hfill (1)

In Eq. 1, $E$ represents the elasticity modulus, $\alpha$ represents the coefficient of thermal expansion of the material of the beam, $\sigma_x$ represents the component of normal stress acting in the $x$ direction and $T$ represents the temperature distribution acting along the width of the beam. In [6] the constants $C_1$ and $C_2$ were found from the hypothesis that the component of stress $\sigma_x$ is self-balanced. This hypothesis was assumed by the fact that there are no external loads acting at the ends of the beam. To be self-balanced, the component of stress $\sigma_x$ should have the result of efforts made by it null in cross-section of the beam, i.e.

$$\int_{-h}^{+h} \sigma_x b dy = \int_{-h}^{+h} \sigma_x d y = 0$$  \hfill (2)

$$\int_{-h}^{+h} \sigma_x b y dy = \int_{-h}^{+h} \sigma_x y dy = 0$$  \hfill (3)

In [5] it was assumed that the plate of FIG. 1 was subjected to a generalized polynomial temperature distribution as follows:

$$T(y) = a_0 + a_1 y + a_2 y^2 + \cdots + a_n y^n$$  \hfill (4)

The temperature distribution given by Eq. 4 represents the polynomial approximation for the nonlinear temperature field acting on the plate.

In the Eq. 4 the terms $a_0$, $a_1$, $a_2$, ..., $a_n$ represent the polynomial coefficients of the approximate temperature distribution. Thus, the work presented in [5] was replaced the Eq. 4 in Eq. 1. The result obtained was replaced in Eq. 2 and Eq. 3 to find the component of stress $\sigma_x$ in terms of the polynomial coefficients of the approximate temperature distribution, i.e.

$$\sigma_x = E\alpha \sum_{n=0}^{N} \left( \frac{a_{2n} h^{2n}}{2n+1} + \frac{3a_{2n+1} h^{2n} y}{2n+3} - a_n y^n \right)$$  \hfill (5)

In possession of the component of stress $\sigma_x$, in [5] it was shown that the components of strain $\varepsilon_x$, $\varepsilon_y$ and $\varepsilon_z$ acting on the plate are given by:

$$\varepsilon_x = \alpha \sum_{n=0}^{N} \left( \frac{a_{2n} h^{2n}}{2n+1} + \frac{3a_{2n+1} h^{2n} y}{2n+3} \right)$$  \hfill (6)

$$\varepsilon_y = -\nu\alpha \sum_{n=0}^{N} \left( \frac{a_{2n} h^{2n}}{2n+1} + \frac{3a_{2n+1} h^{2n} y}{2n+3} \right) + \alpha(y+1) \sum_{n=0}^{N} (a_n y^n)$$  \hfill (7)

$$\varepsilon_z = -\nu\alpha \sum_{n=0}^{N} \left( \frac{a_{2n} h^{2n}}{2n+1} + \frac{3a_{2n+1} h^{2n} y}{2n+3} \right) + \alpha(y+1) \sum_{n=0}^{N} (a_n y^n)$$  \hfill (8)

In the Eq. 5, Eq. 6, Eq. 7 and Eq. 8, $2n$ represents the even terms, $2n+1$ represents the odd
terms and $h$ represents the half of the width of the plate, and in Eq. 7 and Eq. 8, $\nu$ represents the Poisson’s ratio of the material of the plate.

It is worth noting that, according to the Saint-Venant principle, the solutions given by Eq. 5, Eq. 6, Eq. 7 and Eq. 8 are valid for points located at distances greater than or equal to the width of the plate from the ends that are intercepted by the $x$ axis.

3 CASE STUDY

3.1 Plate in the forced cooling process

In order to exemplify the use of Eq. 5, Eq. 6, Eq. 7 and Eq. 8, it was considered the case where a plate was subjected to a temperature distribution along its width. In this way, it was considered the temperature distribution on a plate that has been subjected to the forced cooling process. This process is shown in FIG. 2.

The plate, before going through the process of forced cooling, was submitted to the process of hot rolling. Thus, the plate out of the hot rolling process at a temperature $T_0$. It was assumed that the heat exchange of the plate with the environment occurred only within the cooling unit and that the width of the cooling unit was equal to the width of the plate, i.e., in a moment of time the plate was completely immersed in the cooling unit. The temperature distribution on the plate was found at the instant that it was fully immersed in the cooling unit. The temperature distribution on the plate was found by [8] and is given by:

$$T(y) = S_1 e^{r_1 y} + S_2 e^{r_2 y} + K \quad \text{(9)}$$

Where $r_1$ and $r_2$ are given by:
The value of $K$ is given by:

$$K = \frac{T_\infty h^*}{kb} \left( \frac{1}{r_2} - \frac{1}{r_1} \right)$$

(12)

And the values of $K$ is.

The descriptions of the parameters that appear in the process, as well as the values for these parameters, are shown through the TAB. 1. It was considered that the material of the plate was the carbon steel SAE 1020.

<table>
<thead>
<tr>
<th>Description</th>
<th>Symbol</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Width of the Plate</td>
<td>$2h$</td>
<td>0.5 m</td>
</tr>
<tr>
<td>Length of the Plate</td>
<td>$2l$</td>
<td>3 m</td>
</tr>
<tr>
<td>Cooling Unit Temperature</td>
<td>$T_\infty$</td>
<td>25 $^\circ$C</td>
</tr>
<tr>
<td>Temperature at $y=-h$</td>
<td>$T_0$</td>
<td>100 $^\circ$C</td>
</tr>
<tr>
<td>Thermal Conductivity</td>
<td>$k$</td>
<td>51.9 W/mK</td>
</tr>
<tr>
<td>Plate Thickness</td>
<td>$2b$</td>
<td>0.01 m</td>
</tr>
<tr>
<td>Thermal Diffusivity</td>
<td>$a$</td>
<td>$1.357 \times 10^{-5}$ m$^2$/s</td>
</tr>
<tr>
<td>Coefficient of Heat Exchange</td>
<td>$h^*$</td>
<td>1000 W/mK</td>
</tr>
<tr>
<td>Speed of Lamination</td>
<td>$u$</td>
<td>0.01 m/s</td>
</tr>
</tbody>
</table>

The constants $S_1$ and $S_2$ of Eq. 9 were found by means of the following boundary conditions:

$$T(y)|_{y=-h} = T_0$$

(13)

$$\frac{dT(y)}{dy} = 0, y = h$$

(14)

Substituting the values of the parameters shown in TAB. 1 in Eq. 10, Eq. 11 and Eq. 12 and applying the boundary conditions given by Eq. 13 and Eq. 14, it was found the following temperature distribution on the plate:

$$T(y) = 9.867 \times 10^{-83} e^{7.423 \times 10^2 y} + 20.485 e^{-5.191 y} + 25$$

(15)

### 3.2 Polynomial approximation for the given temperature profile

The temperature profile given by Eq. 15 was approximated by a polynomial to enable the use of the relations presented previously on section 2 in order to determine the stress and strain fields acting on the plate of FIG. 2. A linear combination of the first Legendre polynomials was used and its coefficients determined by means of the Least Squares method. The Legendre polynomials form a family of orthogonal polynomials, which can be
characterized as an advantage in computer terms in the application of the method of Least Squares and, moreover, the family of Legendre polynomials has a constant and unitary weight function, which will account for equal weight in the approximation for the entire interval. In the application of the Least Squares method the function $T(y)$, given by Eq. 15, was approximated by a function $g(y)$ given by:

$$g(y) = c_1\phi_1 + c_2\phi_2 + c_3\phi_3 + c_4\phi_4 + c_5\phi_5 + c_6\phi_6 + c_7\phi_7 + c_8\phi_8 + c_9\phi_9$$

(16)

In Eq. 16, the first nine Legendre polynomials were used for the approximation, moreover the use of a number greater than this would represent high numerical processing time without accounting for a large gain in accuracy. In this equation, the constants coefficients $c_1, c_2, \ldots, c_9$ were determined by the Least Squares methods and the functions $\phi_1, \phi_2, \ldots, \phi_9$ represents the first nine Legendre polynomials which are shown in TAB. 2.

**Table 2: First nine Legendre polynomials**

<table>
<thead>
<tr>
<th>Function $\phi_i$</th>
<th>Polynomial</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi_1$</td>
<td>1</td>
</tr>
<tr>
<td>$\phi_2$</td>
<td>y</td>
</tr>
<tr>
<td>$\phi_3$</td>
<td>1/2(3y^2-1)</td>
</tr>
<tr>
<td>$\phi_4$</td>
<td>1/2(5y^3-3y)</td>
</tr>
<tr>
<td>$\phi_5$</td>
<td>1/8(35y^3-30y^2+3)</td>
</tr>
<tr>
<td>$\phi_6$</td>
<td>1/8(63y^4-70y^3+15y)</td>
</tr>
<tr>
<td>$\phi_7$</td>
<td>1/16(231y^5-315y^4+105y^2-5)</td>
</tr>
<tr>
<td>$\phi_8$</td>
<td>1/16(429y^6-693y^5+315y^3-35y)</td>
</tr>
<tr>
<td>$\phi_9$</td>
<td>1/128(6435y^7-12012y^6+6930y^4-1260y^2+35)</td>
</tr>
</tbody>
</table>

The application of the Least Squares procedure consists in minimizing the error in the interval of approximation. The mean square error in the range of approximation is given by:

$$\int_{-h}^{+h} r^2(y) = \int_{-h}^{+h} [T(y) - g(y)]^2 dy$$

(17)
The values for the constants \( c_1, c_2, \ldots, c_9 \) were found by replacing Eq. 15 and Eq. 16 in Eq. 17 and then finding the minimum mean square error. Replacing the constants found in Eq. 16 and performing the necessary simplifications, it was found the following polynomial approximation for the temperature distribution:

\[
g(y) = 45.486 - 106.349y + 275.871y^2 - 476.398y^3 + 635.136y^4 - 694.269y^5 + 61.382y^6 + 163.321y^7 + 5140.984y^8
\]

(18)

The original temperature distribution, given by Eq. 15, as well as the approximate temperature distribution, given by Eq. 18, are represented in the same graph of FIG. 3.

3.3 Approximate analytical solution for the stress and strain fields

The stress and strain fields acting on the plate were determined using the polynomial coefficients of the approximate temperature distribution given by Eq. 18 and through the mechanical properties of the carbon steel SAE 1020 that are given by TAB. 3.

<table>
<thead>
<tr>
<th>Property</th>
<th>Symbol</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elasticity Modulus</td>
<td>( E )</td>
<td>200x10^9 Pa</td>
</tr>
<tr>
<td>Poisson’s Ratio</td>
<td>( \nu )</td>
<td>0.3</td>
</tr>
<tr>
<td>Coefficient of Thermal Expansion</td>
<td>( \alpha )</td>
<td>11.9x10^-6 C^-1</td>
</tr>
</tbody>
</table>

The polynomial coefficients of the approximate temperature distribution are given by TAB. 4.

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a_0 )</td>
<td>45.486</td>
</tr>
<tr>
<td>( a_1 )</td>
<td>-106.349</td>
</tr>
<tr>
<td>( a_2 )</td>
<td>275.871</td>
</tr>
<tr>
<td>( a_3 )</td>
<td>-476.398</td>
</tr>
<tr>
<td>( a_4 )</td>
<td>635.136</td>
</tr>
<tr>
<td>( a_5 )</td>
<td>-694.269</td>
</tr>
<tr>
<td>( a_6 )</td>
<td>61.382</td>
</tr>
<tr>
<td>( a_7 )</td>
<td>163.321</td>
</tr>
<tr>
<td>( a_8 )</td>
<td>5140.984</td>
</tr>
</tbody>
</table>

The replacement of the polynomial coefficients, the value of \( h \) and the mechanical properties of the carbon steel SAE 1020 in Eq. 5, Eq. 6, Eq. 7 and Eq. 8 resulted in the following polynomials for the component of stress \( \sigma_x \), and for the components of strain \( \varepsilon_x \), \( \varepsilon_y \) and \( \varepsilon_z \):

\[
\sigma_x (y) = 14885388.024 - 45253089.152y - 656572408.485y^2 + 1133826542.940y^3 - 1511624462.966y^4 + 1652359405.121y^5 - 146088846.423y^6 - 388704957.83y^7 - 12235541640.283y^8
\]

(19)

\[
\varepsilon_x (y) = 0.000616 - 0.00149y
\]

(20)
3.4 Solution using the finite element method

In this section, the numerical solutions for the stress and strain fields acting on the plate are presented. These solutions were determined by the application of the finite element method using the Ansys program.

The long flat rectangular plate, shown by FIG. 1, was modeled by the Ansys program using the condition of symmetry that it has with respect to the y-axis that passes through the origin of the coordinate system. Thus only the right half of the plate was modeled. In the application of the method were used 120000 elements. The element used in the analysis was a plane element designated by Plane 183. The model used is shown in FIG. 4, as well as the segments where numerical data were collected.

In the plate modeling, the dimensions provided by TAB. 1 were used, i.e., $2h=0.5m$ and $2l=3m$.

$$\varepsilon_y(y) = 0.000519 - 0.00120y + 0.00427y^2 - 0.00737y^3 + 0.00983y^4 - 0.0107y^5 + 0.000950y^6 + 0.00253y^7 + 0.0795y^8$$  \hfill (21)

$$\varepsilon_x(y) = 0.000519 - 0.00120y + 0.00427y^2 - 0.00737y^3 + 0.00983y^4 - 0.0107y^5 + 0.000950y^6 + 0.00253y^7 + 0.0795y^8$$  \hfill (22)

3.5 Comparison of numerical and approximate analytical solutions

In the model shown in FIG. 4 was imposed the distribution of temperature given by Eq. 15. The numerical solutions for the component of stress $\sigma_x$, as well as for the components of strain $\varepsilon_x$, $\varepsilon_y$ and $\varepsilon_z$, were evaluated at different distances from the right free end of the plate. The solutions were evaluated in the segments defined by $x=0.0m$, $x=0.5m$, $x=1.0m$ and $x=1.25m$. These segments are shown by FIG. 4.

Being the width of the plate equal to $2h=0.5m$ it can be expected, in accordance with the Saint-Venant principle that, the solutions given by Eq. 19, Eq. 20, Eq. 21 and Eq. 22 are satisfactory for points located at distances greater than or equal to $2h=0.5m$ of the right free
end of the plate. In this way, the numerical solutions for each segment were compared with the approximate analytical solutions given by Eq. 19, Eq. 20, Eq. 21 and Eq. 22. The comparative results for the component of stress $\sigma_x$ and for the components of strain $\varepsilon_x$, $\varepsilon_y$ and $\varepsilon_z$, are shown by means of FIG. 5, FIG. 6, FIG. 7 and FIG. 8, respectively.

**Figure 5**: Results for the component of stress $\sigma_x$

**Figure 6**: Results for the component of strain $\varepsilon_x$
3.6 Presented disagreements between the solutions

The disagreements between the solutions were evaluated at the points where the components of stress and strain presented their maximum and minimum values for the approximate analytical solution. The disagreements are presented for the component of stress $\sigma_x$ and for the components of strain $\varepsilon_{xy}$, $\varepsilon_y$, and $\varepsilon_z$ through the TAB. 5, TAB. 6, TAB. 7 and TAB. 8, respectively.

<table>
<thead>
<tr>
<th>Table 5: Percentage disagreement for the component of stress $\sigma_x$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Point</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>Maximum</td>
</tr>
<tr>
<td>Minimum</td>
</tr>
</tbody>
</table>
Table 6: Percentage disagreement for the component of strain $\varepsilon_x$

<table>
<thead>
<tr>
<th>Point</th>
<th>Y-Coordinate</th>
<th>x=0.00m</th>
<th>x=0.50m</th>
<th>x=1.00m</th>
<th>x=1.25m</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum</td>
<td>-0.25</td>
<td>0.00021%</td>
<td>0.00485%</td>
<td>0.0407%</td>
<td>3.231%</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.25</td>
<td>0.00055%</td>
<td>0.0211%</td>
<td>0.177%</td>
<td>12.065%</td>
</tr>
</tbody>
</table>

Table 7: Percentage disagreement for the component of strain $\varepsilon_y$

<table>
<thead>
<tr>
<th>Point</th>
<th>Y-Coordinate</th>
<th>x=0.00m</th>
<th>x=0.50m</th>
<th>x=1.00m</th>
<th>x=1.25m</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum</td>
<td>-0.25</td>
<td>0.00115%</td>
<td>0.00115%</td>
<td>0.00914%</td>
<td>0.769%</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.245</td>
<td>0.0202%</td>
<td>0.0152%</td>
<td>0.0601%</td>
<td>2.159%</td>
</tr>
</tbody>
</table>

Table 8: Percentage disagreement for the component of strain $\varepsilon_z$

<table>
<thead>
<tr>
<th>Point</th>
<th>Y-Coordinate</th>
<th>x=0.00m</th>
<th>x=0.50m</th>
<th>x=1.00m</th>
<th>x=1.25m</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum</td>
<td>-0.25</td>
<td>0.00115%</td>
<td>0.00115%</td>
<td>0.00914%</td>
<td>0.769%</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.245</td>
<td>0.0202%</td>
<td>0.0152%</td>
<td>0.0626%</td>
<td>2.167%</td>
</tr>
</tbody>
</table>

4 DISCUSSION OF RESULTS

4.1 Component of stress $\sigma_x$

Through the FIG. 5 it was observed that only the numerical solution at $x=1.25m$ did not provide a satisfactory approximation to the approximate analytical solution. This segment is outside the limit of validity of the solution, because it is at a distance less than $2h=0.5m$ of the right free end of the plate. Observing the discrepancies presented in TAB. 5 it can be verified that only in this segment the disagreements were relevant, being as large as 28.484% at the point of maximum stress and 15.853% at the point of minimum stress. For the other segments the discrepancies have not reached 2%, both for the point of maximum stress as well as to the point of minimum stress. In this way, it was determined that the approximate analytical solution to the component of stress $\sigma_x$, given by Eq. 19, was satisfactory for the points located at distance greater than or equal to $2h=0.5m$ of the free end of the plate.

4.2 Components of strain $\varepsilon_{xx}$, $\varepsilon_y$ and $\varepsilon_z$

For the components of strain $\varepsilon_{xx}$, $\varepsilon_y$ and $\varepsilon_z$, it was possible to draw conclusions similar to those obtained for the component of stress $\sigma_x$. With respect to the components of strain $\varepsilon_x$ and $\varepsilon_y$, it can be observed through the FIG. 6 and FIG. 7 that, only the numerical solution at $x=1.25m$ did not provide a satisfactory approximation to the approximate analytical solutions. Observing the discrepancies presented in TAB. 6 and TAB. 7 it can be verified that only in this segment the discrepancies were relevant. For this segment, the largest discrepancy was 3.231% at the point of maximum strain and 12.065% at the point of minimum strain for the component of strain $\varepsilon_x$. For the component of strain $\varepsilon_y$ the discrepancies were 0.769% at the point of maximum strain and 2.159% at the point of minimum strain. For the other segments the disagreements have not reached 1%, both for the point of maximum strain as well as to the point of minimum strain with respect to the two components of strain. In this way, it was determined that the approximate analytical solutions to the components of strain $\varepsilon_x$ and $\varepsilon_y$, given by Eq. 20 and Eq.
21, were satisfactory for points situated at distance greater than or equal to $2h=0.5m$ of the free end of the plate.

For the component of strain $\varepsilon_z$, it can be observed in the FIG. 8 that the numerical solution at $x=1.25m$ has a more satisfactory approximation to the approximate analytical solution than the one presented for the components of strain $\varepsilon_y$. The disagreement presented in this segment was the same as that presented for the component of strain $\varepsilon_y$ at the point of maximum strain, i.e., it was $0.769\%$. With regard to the point of minimum strain, the disagreement was a little higher than the one presented for the component of strain $\varepsilon_y$, being of $2.167\%$. In this way, it was determined that the approximate analytical solution for the component of strain $\varepsilon_z$, given by Eq. 22, was satisfactory for the points located at distances greater than or equal to $2h=0.5m$ of the free end of the plate.

5 CONCLUSION

The approximate analytical results showed a high level of agreement with the numerical results at points away from the free ends that are intercepted by the $x$ axis. The high level of accuracy in the results shows that the expressions presented by Eq. 5, Eq. 6, Eq. 7 and Eq. 8 represent in a satisfactory way the stress and strain fields acting in long flat rectangular plates that are isotropic and free of external loads at their ends. These expressions are satisfactory for points located at distances greater than or equal to the width from the free ends that are intercepted by the $x$ axis. In this way, it is determined a formulation for the rapid and efficient determination of self-balanced thermal stresses and thermal strains in structures under adverse temperature variations that is hardly found in technical literature.

6 REFERENCES


NUMERICAL SIMULATION FOR IDENTIFICATION OF FREE SPANS IN SUBMARINE PIPELINES

BRUNO M. JACOVAZZO\textsuperscript{1,2}, BRUNO F. MONTEIRO\textsuperscript{1,2}, MAURO H. A. DE LIMA JUNIOR\textsuperscript{1}, JULIANA S. BAIIOCO\textsuperscript{1,3}, FABRÍCIO N. CORRÊA\textsuperscript{1,2}, BRENNO P. JACOB\textsuperscript{1}

\textsuperscript{1} LAMCOSO/COPPE/UFRJ Rio de Janeiro, RJ, Brazil.
\textsuperscript{2} POLI/UFRJ Rio de Janeiro, RJ, Brazil.
\textsuperscript{3} TEQ/ E. Engenharia/UFF, Niterói, RJ, Brazil.
LAMCOSO—Laboratory of Computational Methods and Offshore Systems, Departamento de Engenharia Civil, Centro de Tecnologia, Bloco B sala B-101, Cidade Universitária, Ilha do Fundão, Caixa Postal 68.506, 21941-914 Rio de Janeiro, RJ, Brasil.
\{bruno, brunomonteiro, mauro, jsbaioco, fabricio, breno\}@lamcso.coppe.ufrj.br / http://www.lamcso.coppe.ufrj.br

{brunojacovazzo, bruno.monteiro, fabricio}@poli.ufrj.br / http://www.poli.ufrj.br/

jsbaioco@id.uff.br / http://www.engenharia.uff.br/

Key words: Pipe Laying, Free Span, Vortex Induced Vibration (VIV).

Abstract. It is well known that, after a submarine pipeline is laid over the seabed, free spans may appear due to the unevenness of the seafloor, associated to the relatively high bending stiffness of the pipe, and to the residual tension from the pipe laying process.

The identification of free spans along a given pipeline route is very important due to several factors. One of them is to allow the verification of design criteria related to fatigue induced by vortex induced vibrations (VIV), such as those presented on the DNV-RP-F105 code [1]. Also, the number and length of free spans on a route may require extensive intervention works (e.g. rock dumping, digging, etc) to mitigate the spans and reduce unacceptable span lengths; or, alternatively, may require changes in the predetermined route of the pipeline.

In this context, a numerical tool has been developed for the determination of the final equilibrium configuration of a pipeline after the execution of its laying procedure [2], [3]. The tool is based on global Finite Element models, and simulates the evolution of the pipe laying process along a predetermined route, using the actual seafloor bathymetry data. The objective is to allow an easier implementation of the verification of VIV induced fatigue, and the assessment of a given route in terms of the number and length of its free spans; for this purpose the tool may be associated to a program oriented towards the synthesis and
optimization of submarine pipelines.

In this work, a case study is presented to illustrate the use of the tool. In this case, a route is proposed to connect the pipeline from one location to another over the seabottom, represented by a complex bathymetry of the soil and by its physical properties.

1 INTRODUCTION

Submarine pipelines are mainly used to transport the production of offshore platforms. Their installation are performed by several methods [2] in which the pipeline is laid over the seabed and connected from one point to another. The definition of submarine pipeline routes is a complex task that involves many variables and requires sensibility and experience from the designer. To minimize costs, it would be ideal to establish a straight route between the two connection points of the pipeline; however in general this is not feasible due to several problems that can appear in the route.

One of them is related to the formation of free spans. After laying the pipeline on the seabed, the unevenness of the seafloor associated to the relatively high bending stiffness of the pipe may lead to the formation of free spans along the route. Figure 1 schematically presents an example of free span.

The current flow acting on the suspended portion of the pipe may generate vortex shedding (as represented in Figure 2) which induces vibrations on the pipeline. This phenomenon is known as “Vortex Induced Vibration” (VIV) and may generate fatigue problems at the pipeline.

![Figure 1: Schematic example of a free span](image1)

![Figure 2: Vortex shedding in a pipeline with the flow action](image2)
the length of a pipeline route (in order to avoid problems like the one previously described) and the cost of engineering interventions that may be required to solve problems of routes with shorter length.

2 OBJECTIVE

Considering the problem presented above, engineers should be able to identify where free spans may occur when designing a pipeline route in order to verify if there would be fatigue problems at the pipe. However this is not a simple task as the formation of free spans depends on several factors inherent to the installation procedure and to the pipeline and soil characteristics. Some of them can be listed below:

- Launching tension applied to the pipeline by the vessel.
- Bathymetry data, giving the soil profile for a determined route.
- Soil data as the vertical stiffness and friction properties.
- Pipe section data that would give information about weight, bending stiffness and more of the pipeline.

In this context, a numerical tool has been developed to obtain the final equilibrium configuration of submarine pipelines on the seabed after the laying operation, by generating Finite Element models [4],[5],[6] and performing nonlinear analyses while automatically activating/deactivating finite elements along the evolution of the pipe laying operation. Thus, this numerical tool is able to provide several details of the final configuration of the pipeline after the installation process, such as its stresses and free spans characteristics. Then, the engineer would be able to apply criteria based on the recommendations of DNV-RP-F105 code in order to investigate possible fatigue problems due to VIV at any given candidate route.

The numerical tool has been implemented in the SITUA-Prosim system [7],[8], that has been developed by LAMCSO-COPPE-UFRJ in collaboration with CENPES-Petrobras. Originally, this system was oriented to the analysis of moored floating units considering its coupling with the mooring lines and risers; currently it includes several special features for analyzing different installation and failure situations (including pipeline installation).

The SITUA platform incorporates some specific features to the generation of models for the simulation of installation procedures for offshore pipelines [9],[10]; specifically, the numerical tool presented in this work has been incorporated on the PETROPIPE module [11], which allows an accurate representation of the vessel devices that support the pipeline during the launching process, including the ramp and stinger rollers, representing the contact between these elements and the pipeline.

The details of the numerical tool development have been presented in [3], where a simple case study has been performed. The main objective of this paper is to show the application of the numerical tool considering a more complex case study, with real bathymetry data that could be representative of a real project.

3 NUMERICAL TOOL

As described above, details of the numerical tool development have been presented in [3]. However, it is summarized below.

Figure 3a schematically presents a pipeline being launched by a vessel. As the objective of
the numerical tool is to obtain the final equilibrium configuration of submarine pipelines on the seabed after the laying operation and assuming that the design of the pipe laying procedure itself will include solutions in order to maintain the pipeline being launched along the predetermined route, considering an acceptable environmental window of operation, the action of environmental loads is not considered in the model.

For the same reason that has been described above, as the behavior of the suspended portion of the pipeline during the laying operation will not be investigated, the example that has been presented in Figure 3a can be truncated as shown in Figure 3b. For this configuration, the vertical component acting on the truncation point is equal to the submerged weight of the remaining suspended pipeline segment. The horizontal component must match the design value determined for the tensioner system during the installation operation, so that the pipe is able to maintain a minimum tolerable radius of curvature at the TDP region. This force must be controlled by the algorithm implemented in the program.

Figure 3: Truncation of the model

The simulation of the progressive pipe laying process is performed by prescribing incremental displacements at the truncation point of the model following the predetermined route, at every step of successive static analyses (quasi-static simulation), as can be seen in Figure 4. As the launching point progresses along the route, the horizontal component of the force acting at the truncation point would increase. Therefore, a procedure to control the value of this force during the simulation, and maintain it close to the design value defined by the user has been implemented on the numerical tool.

Figure 4: Pipeline configuration after applying a prescribed displacement at the launching point
On each step of the quasi-static simulation, the program verifies the value of the horizontal force component, and compares it with the specified design value. Whenever the horizontal force obtained in a given simulation step is higher than the design value, the program automatically increases the undeformed length of the top element of the mesh so that the force returns to the specified value. Thus, the total length of the pipeline is progressively increased, leading it to accommodate on the seabed as the pipeline is launched.

However, if the top element length were increased indefinitely, the mesh discretization would be impaired, leading to instability and convergence problems in the solution method of the nonlinear problem, reducing the quality of the results. To solve this problem, the simulation procedure automatically subdivides the top element whenever its length exceeds twice the original length. This leads to the introduction of new finite elements in the mesh.

4 CASE STUDY

This case study presents the execution of two pipe laying simulations on a route with irregular bathymetry and two types of pipelines, with different sections.

3.1 Bathymetry and pipeline route

The adopted bathymetric sea bottom to perform this case study, shown in Figure 5, is a rectangular cutout of a real micro-bathymetry data, with dimensions equal to 4160m x 3140m and adjacent contour lines representing variations of depth equal to 1.5m. The global reference axis is located at the sea still water level at the center of the rectangular cutout; the average depth of the region is about 800m. During the execution of the pipe laying simulations, the bathymetric sea bottom has been represented by a rectangular mesh containing elements measuring 10m each.

The defined route for this case study is presented in Figure 6. In this figure the locations of the global axis and the initial and final points of the pipeline route ("A" and "B" respectively) are represented. The exact locations of these points related to the global axis are described in Table 1.

The route begins with a straight line of 900m towards the global X-axis. Then, a 500m radius curve is described along an angle of 73.6°. Next, the route follows a straight line of 1000m, where it reaches a very rugged region. Then, a new 500m curve is described along an angle of 73.6°. Finally, the route follows another straight line of 1100m in the direction of the global X-axis reaching the point "B".

Certainly, this is not the best route to connect the point "A" to point "B". However, it was defined this way as an example just because it crosses a very rugged region, capable of generating free spans that should be identified by the algorithm.
Figure 5: Irregular bathymetry – 3D view

Figure 6: Defined pipeline route
3.2 Soil data

The physical characteristics of the soil are represented by the same model traditionally used in programs such as Prosim and Anflex [7], [12]: the pipe-soil contact is represented by vertical linear springs; the friction between the soil and the pipe is represented by an anisotropic elastoplastic model in which distinct friction coefficients are defined considering axial and transversal directions of the pipe. In this case, a value of 300kN/m/m has been adopted for the stiffness of the vertical spring.

3.3 Pipe section

Two different pipelines have been simulated along the proposed route: with 8 and 16 inches of nominal diameter. The objective of this approach was to evaluate the differences in the results considering a more flexible and a more rigid pipe. Table 2 presents the main characteristics of the pipe sections.

A 1m length mesh has been used to represent the pipelines. Considering that the micro-bathymetry used in this case study has adjacent contour lines representing variations of depth equal to 1.5m, the mesh discretization of the pipelines is adequate.

The horizontal tension applied at the truncation point has been defined with the purpose to maintain the stresses in the pipelines below a limit of 80% [13] of the yield stress of the steel during the installation procedure at the TDP. Thus, the horizontal tensions have been defined as equal to 57,5kN and 190kN for the 8” and 16” pipelines, respectively. It should be noted that when the pipeline is accommodated over the seabed, it assumes new curvatures that depends on many characteristics including the soil profile. This can lead the pipes to have stress levels that exceed the limit of 80% of the steel yield stress.

Table 2: Main characteristics of the pipes

<table>
<thead>
<tr>
<th>Property</th>
<th>8 inches</th>
<th>16 inches</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>External diameter</td>
<td>8.625</td>
<td>16.000</td>
<td>in</td>
</tr>
<tr>
<td>Wall thickness</td>
<td>0.50</td>
<td>0.75</td>
<td>in</td>
</tr>
<tr>
<td>Steel Young Modulus</td>
<td>207</td>
<td>207</td>
<td>GPa</td>
</tr>
<tr>
<td>Steel yield stress (X60 [14])</td>
<td>414</td>
<td>414</td>
<td>MPa</td>
</tr>
<tr>
<td>Steel specific weight</td>
<td>77.0</td>
<td>77.0</td>
<td>kN/m³</td>
</tr>
<tr>
<td>Coating thickness</td>
<td>2.00</td>
<td>2.00</td>
<td>in</td>
</tr>
<tr>
<td>Coating specific weight (concrete)</td>
<td>23.5</td>
<td>23.5</td>
<td>kN/m³</td>
</tr>
</tbody>
</table>
3.4 Results

**Deformed configuration**

Figures 7 and 8 present a tridimensional view of the final configuration of the pipelines after the execution of the pipe laying procedure.

*Figure 7: Final configuration of the pipeline laying on the seabed – 8 inches pipe*
Figure 8: Final configuration of the pipeline laying on the seabed – 16 inches pipe
**Free spans identification**

Table 3 presents the location and the length of the free spans that have been identified by the numerical algorithm after the execution of the simulations.

<table>
<thead>
<tr>
<th>Free Span</th>
<th>Initial Point (m)</th>
<th>Final Point (m)</th>
<th>Length (m)</th>
<th>Free Span</th>
<th>Initial Point (m)</th>
<th>Final Point (m)</th>
<th>Length (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1661</td>
<td>1689</td>
<td>28</td>
<td>1</td>
<td>956</td>
<td>968</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>1698</td>
<td>1774</td>
<td>76</td>
<td>2</td>
<td>1112</td>
<td>1121</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>1849</td>
<td>1863</td>
<td>14</td>
<td>3</td>
<td>1610</td>
<td>1771</td>
<td>161</td>
</tr>
<tr>
<td>4</td>
<td>1966</td>
<td>2027</td>
<td>61</td>
<td>4</td>
<td>1817</td>
<td>1891</td>
<td>74</td>
</tr>
<tr>
<td>5</td>
<td>2050</td>
<td>2136</td>
<td>86</td>
<td>5</td>
<td>1914</td>
<td>2076</td>
<td>162</td>
</tr>
<tr>
<td>6</td>
<td>2294</td>
<td>2325</td>
<td>31</td>
<td>6</td>
<td>2087</td>
<td>2165</td>
<td>78</td>
</tr>
<tr>
<td>7</td>
<td>2357</td>
<td>2398</td>
<td>41</td>
<td>7</td>
<td>2262</td>
<td>2308</td>
<td>46</td>
</tr>
<tr>
<td>8</td>
<td>2432</td>
<td>2456</td>
<td>24</td>
<td>8</td>
<td>2323</td>
<td>2438</td>
<td>115</td>
</tr>
<tr>
<td>9</td>
<td>2640</td>
<td>2705</td>
<td>65</td>
<td>9</td>
<td>2464</td>
<td>2488</td>
<td>24</td>
</tr>
<tr>
<td>10</td>
<td>3070</td>
<td>3086</td>
<td>16</td>
<td>10</td>
<td>2522</td>
<td>2546</td>
<td>24</td>
</tr>
<tr>
<td>11</td>
<td>3600</td>
<td>3609</td>
<td>9</td>
<td>11</td>
<td>2561</td>
<td>2591</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>12</td>
<td>2602</td>
<td>2692</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13</td>
<td>3013</td>
<td>3050</td>
<td>37</td>
</tr>
</tbody>
</table>

**8 inches pipeline**

**16 inches pipeline**

**Pipeline stresses**

Figures 9 and 10 present the residual axial tensions and the Von Mises stresses, respectively, along the pipeline length laying on the seabed after the execution of the simulations.

**Figure 9**: Residual tensions along the pipelines after the execution of the simulations
2 CONCLUSIONS

As expected, it can be observed that the 8 inches pipeline has better accommodated on the seabed due to its lower bending stiffness value in comparison with the 16 inches pipeline. Therefore, a higher number of free spans have been identified by the algorithm for the 16 inches pipeline (Table 3).

The residual tensions along the pipelines (that are maintained by the friction force of the soil - Figure 9) have been higher for the 16 inches pipeline as the horizontal tension applied at the truncation point of the model was higher for this pipeline.

As it can be observed in Figure 10, both pipelines have violated the yield stress limit of the steel [13] in some points of the route, primarily at its most rugged section (showed in Figures 7 and 8). The major peaks of stress could be lower if the soil had been represented by an even more refined mesh.

2 ACKNOWLEDGMENTS

The authors would like to thank the support of the National Council for Scientific and Technological Development (CNPq) and the Carlos Chagas Filho Foundation for Research Support of the Rio de Janeiro State (FAPERJ).

The authors would also like to acknowledge the active support of the Research & Development Center of PETROBRAS (CENPES).

REFERENCES


Figure 10: Von Mises stresses along the pipelines after the execution of the simulations


ON THE SEISMIC RESPONSE OF FRACTURES AND INDUCED ANISOTROPY IN FLUID-SATURATED POROUS MEDIA

ROBIEEL MARTÍNEZ CORREDOR1, JUAN E. SANTOS2, PATRICIA M. GAUZELLINO3 AND JOSÉ M. CARCIONE4

1Facultad de Ingeniería, Universidad Nacional de La Plata
Calle 1 esquina 47, La Plata, Argentina
robielmartinez@yahoo.com

2Instituto del Gas y del Petróleo, Facultad de Ingeniería, Universidad de Buenos Aires y Universidad Nacional de La Plata. Department of Mathematics, Purdue University, USA.
Las Heras 2214 Piso 3 C1127AAR Buenos Aires, Argentina
santos@math.purdue.edu

3Facultad de Cs. Astronómicas y Geofísica, Universidad Nacional de La Plata
Paseo del Bosque s/n, La Plata, Argentina
gauze@fcaglp.unlp.edu.ar

4Istituto Nazionale di Oceanografia e di Geofisica Sperimentale (OGS)
Borgo Grotta Gigante 42c, 34010 Sgonico, Trieste, Italy
jcarcione@inogs.it

Key words: Porous Media, Fractures, Anisotropy, FEM

Abstract. A planar fracture embedded in a fluid-saturated poroelastic - Biot - medium can be modeled either as a extremely thin, highly permeable and compliant porous layer or employing suitable boundary conditions. First we analyze the seismic response at a single fracture separating two poroelastic half spaces. Then the macroscale seismic response of an heterogeneous Biot medium containing a dense set of aligned fractures is determined using a Numerical Rock Physics approach. Finally, we illustrate the propagation of waves at the macroscale for the case of horizontal and vertical aligned fractures employing effective vis coleastic properties determined using our computational rock physics.
1 INTRODUCTION

The study of the wave propagation in fractured porous media is very important in geophysical prospecting. In this work, we start by analyzing the reflection and transmission coefficients for plane waves arriving at a fracture within two poroelastic half spaces. The fracture is modeled either as a highly permeable and compliant thin poroelastic layer or a boundary condition separating the two half spaces. When the fracture is modeled as a thin layer, the boundary conditions at the interfaces separating the layer from the half spaces are the continuity of the displacements and generalized forces, as those given by Dutta and Odé [1]. The plane waves are represented by potentials, from which we derive expressions for the displacement vectors and the stress tensors [2, 3]. The coefficients obtained by this approach are compared with those computed when the fracture is modeled as a boundary condition between the two half spaces [4]. These boundary conditions are given by finite jumps expressed as averaged quantities for displacements, stress and fluid pressure.

Then, we use Numerical Rock Physics in order to obtain effective physical properties of porous media containing a dense set of aligned fractures. Under simplifying hypothesis, it has been shown that this medium behaves as an effective transversely isotropic and viscoelastic (TIV) medium. Our approach allows to determine the complex TIV elastic coefficients for the general case of highly heterogeneous saturant fluids and fracture petrophysical properties. P-waves and S-waves seismic waves traveling in this type of medium show frequency and angular variations of velocity and attenuation due to the mesoscopic loss mechanism caused by slow (diffusion) Biot waves generated at mesoscopic-scale heterogeneities.[5, 6].

We apply a set of compressibility and shear harmonic finite-element (FE) experiments on representative samples of fractured highly heterogeneous Biot media to determine the five complex and frequency dependent stiffnesses characterizing the equivalent TIV medium at the macroscale. For a detailed treatment of this procedure, see [7, 8, 9].

Finally, we illustrate the propagation of waves at the macroscale for a VSP (Vertical Seismic Profile) geometry and the case of horizontal and vertical aligned fractures which determine equivalent media with transversely isotropic symmetry. The effective viscoelastic coefficients are obtained using our Numerical Rock Physics approach. The seismograms are computed by a domain decomposition technique combined with a frequency domain FE method. This numerical strategy has already been applied to wave propagation in 2D and 3D media and to different rheological equations [10, 11, 12].

2 REFLECTION AND TRANSMISSION COEFFICIENTS AT A FRACTURE

When the fracture is modeled as a layer of thickness $h$, our fluid-saturated system consists of three media, $\Omega_n$, $n = 1, 2, 3$ with different properties as shown in Figure 1. Let $z = 0$ be the boundary between the half space $\Omega_1$ and $\Omega_2$, and $z = h$ the boundary
between $\Omega_2$ and the half space $\Omega_3$, and consider a type-I compressional plane wave in $\Omega_1$ incident at $z = 0$ with an angle $\theta_{11}$ with respect to the vertical $z$-axis.

The boundary conditions at the interfaces located at $z = 0$ and $z = h$ impose continuity of the solid and fluid displacements ($U_s, U_z$ and $W_3$), continuity of the normal stress ($\sigma_{zz}$) and continuity of the fluid pressure ($P_f$) \cite{7, 1}. Therefore, at $z = 0$ and $z = h$ we impose the conditions

\begin{align}
\dot{U}_x^{(n)} = \dot{U}_x^{(n+1)}, \quad \dot{U}_z^{(n)} = \dot{U}_z^{(n+1)}, \quad \dot{W}_3^{(n)} = \dot{W}_3^{(n+1)}, \quad \ddot{P}_f^{(n)} = \ddot{P}_f^{(n+1)}, \quad W_1^{(n)} = W_2^{(n+1)}, \quad n = 1, 2. \tag{1}
\end{align}

When the fracture is modeled as a boundary condition separating the half spaces $\Omega_1$ and $\Omega_3$, we employ two different boundary conditions derived in \cite{4}. The first set of boundary conditions, corresponding to their equations (52) are:

\begin{align}
\dot{U}_x^{(3)} - \dot{U}_x^{(1)} = (\omega) \eta_T \sigma_{zz}^{(1)}, \quad \ddot{U}_z^{(3)} - \ddot{U}_z^{(1)} = (\omega) \eta_{ND} \left[ \left( 1 - \alpha^{(2)} \right) B (1 - \Pi) \right] \sigma_{zz}^{(1)} - \alpha^{(2)} \frac{\ddot{P}_f^{(3)} - \dot{P}_f^{(1)}}{2} \Pi, \tag{2}
\end{align}

\begin{align}
W_3^{(3)} - W_3^{(1)} = (\omega) \kappa^{(2)} \eta_{ND} \left[ -\sigma_{zz}^{(1)} - \frac{1}{B} \frac{\ddot{P}_f^{(3)} - \dot{P}_f^{(1)}}{2} \right] \Pi, \quad \sigma_{zz}^{(3)} = \sigma_{zz}^{(1)}, \quad \sigma_{zz}^{(3)} = \sigma_{zz}^{(1)}, \quad -P_f^{(3)} - P_f^{(1)} = \frac{\kappa^{(2)} W_3^{(3)} + W_3^{(1)}}{\kappa(\omega)} \Pi. \tag{3}
\end{align}

The second set of boundary conditions, their equations (53), is a simplified form of (2) (their equations (52) ) when the permeability tends to infinity:

\begin{align}
\dot{U}_x^{(3)} - \dot{U}_x^{(1)} = (\omega) \eta_T \sigma_{zz}^{(1)}, \quad \ddot{U}_z^{(3)} - \ddot{U}_z^{(1)} = (\omega) \eta_{ND} \left[ \sigma_{zz}^{(1)} - \alpha^{(2)} \left( -\dot{P}_f^{(1)} \right) \right], \quad W_3^{(3)} - W_3^{(1)} = (\omega) \alpha^{(2)} \eta_{ND} \left[ -\sigma_{zz}^{(1)} + \frac{1}{B} \left( -\dot{P}_f^{(1)} \right) \right], \tag{4}
\end{align}

\begin{align}
\sigma_{zz}^{(3)} = \sigma_{zz}^{(1)}, \quad \sigma_{zz}^{(3)} = \sigma_{zz}^{(1)}, \quad -P_f^{(3)} = -P_f^{(1)}. \tag{5}
\end{align}

where

\begin{align}
H_D^{(2)} = k_m + \frac{4\mu^{(2)}}{3}, \quad \frac{1}{B} = \alpha^{(2)} + \frac{H_D^{(2)}}{\alpha^{(2)} M^{(2)}} - \frac{\chi}{\alpha^{(2)}}, \quad H_D^{(2)} = \frac{\chi}{2\alpha^{(2)} M^{(2)}}, \quad \beta = 1 - \frac{H_D^{(2)} P_f^{(2)}}{2\alpha^{(2)} M^{(2)} P_f^{(2)}}, \quad \Pi = \frac{\tanh \epsilon}{\epsilon}, \quad \epsilon = \frac{\beta}{2} \tag{6}
\end{align}

\begin{align}
\eta_T = h/\mu^{(2)} \quad \text{(shear compliance)}, \quad \eta_{ND} = h/H_D^{(2)} \quad \text{(dry or drained normal compliance)}, \quad \chi(\omega) = \kappa(\omega)/h \quad \text{(membrane permeability)}. \tag{7}
\end{align}

Here $\chi$ is the vertical component of the polarization vector, and $\beta = (2)$ is the horizontal component of the polarization vector of the slow P-wave. The grains and the fluid are...
characterized by their bulk moduli, $K_s$ and $K_f$, being $\eta_f$ the viscosity of the fluid. $K_m$ and $\mu$ are the bulk and shear modulus of the rock matrix, $\alpha = 1 - (K_m/K_s)$ and $M = ((\alpha - \phi)/K_s + \phi/K_f)^{-1}$. The superscripts indicate the medium, (1 and 3 for the half spaces $\Omega_1$ and $\Omega_3$ and 2 for the fracture parameters).

Table 1: Material properties used for the numerical examples.

<table>
<thead>
<tr>
<th>Properties</th>
<th>Matrix</th>
<th>Fracture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Porosity</td>
<td>0.15</td>
<td>0.5</td>
</tr>
<tr>
<td>Solid density (g/cm$^3$)</td>
<td>2.7</td>
<td>2.7</td>
</tr>
<tr>
<td>Fluid density (g/cm$^3$)</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Fluid viscosity (dyn.s/cm$^2$)</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Solid bulk modulus (dyn/cm$^2$)</td>
<td>$36.0\cdot10^{10}$</td>
<td>$36.0\cdot10^{10}$</td>
</tr>
<tr>
<td>Fluid bulk modulus (dyn/cm$^2$)</td>
<td>$2.25\cdot10^{10}$</td>
<td>$2.25\cdot10^{10}$</td>
</tr>
<tr>
<td>Frame bulk modulus (dyn/cm$^2$)</td>
<td>$9.0\cdot10^{10}$</td>
<td>$5.55\cdot10^8$</td>
</tr>
<tr>
<td>Frame shear modulus (dyn/cm$^2$)</td>
<td>$7.0\cdot10^{10}$</td>
<td>$3.33\cdot10^8$</td>
</tr>
<tr>
<td>Permeability (cm$^2$)</td>
<td>$1.0\cdot10^{-9}$</td>
<td>$1.0\cdot10^{-6}$ (case 1)</td>
</tr>
<tr>
<td>Tortuosity</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

2.1 Numerical Results

The numerical examples consider the experiments for finite fracture permeability (Case C1) and infinite fracture permeability (Case C2).

Figure 2 shows the reflection and transmission coefficients of compressional waves of the incident P-wave at the interfaces $z = 0$ and $z = h$, $h=0.1$ cm. A significant difference between the magnitude of the coefficients obtained with the boundary conditions (1) and (2). In Figure 3, the same experiment was conducted but with smaller $h$, is observed here that the magnitudes obtained with the boundary conditions (1) and the system (52) of Nakagawa-Schoenberg are similar. Hence, when the fracture thickness tends to zero, the two approaches have similar behavior. Figure 4 shows the reflection and transmission coefficients for the limiting case in (3). The magnitudes obtained with the boundary conditions (1) and (3) vary as $h$ increases. When $h = 0.001$ cm, the magnitudes obtained with the two approaches are similar.

3 COMPUTATIONAL ROCK PHYSICS

This kind of rock physics offers a alternative to laboratory measurements with the advantages that they are inexpensive, repeatable and essentially free from experimental errors. Moreover, the numerical experiments may be run using alternative models of the rock and fluid properties. A dense set of parallel plane fractures in a fluid-saturated poroe-
the stiffness matrix (\(p_{ij}\) with \(1 \leq i, j \leq 6\)), we solve the 2D Biot equation of motion in the space-frequency domain with boundary conditions representing compressibility and shear harmonic experiments using Finite Element Method.

Let \(u = (u^s(x), u^f(x))\) be the time Fourier transform of the displacement vector of the solid and fluid relative to the solid frame, respectively. Also let \(\sigma_{kl}(u), p_f(u)\) denote the Fourier transform of the total stress and the fluid pressure, respectively. The frequency-domain constitutive relations are

\[
\sigma_{kl}(u) = 2\mu \varepsilon_{kl}(u^s) + \delta_{kl} \left( \lambda_G \nabla \cdot u^s + \alpha M \nabla \cdot u^f \right), \quad p_f(u) = -\alpha M \nabla \cdot u^s - M \nabla \cdot u^f,
\]

being \(\varepsilon_{kl}(u^s)\) the strain tensor of the solid. In addition, \(\lambda_G = K_G - 2\mu/3\), where \(K_G\) is the bulk modulus of the saturated material. In the diffusive range of frequencies the Biot’s equations are

\[
\nabla \cdot \sigma(u) = 0, \quad i\omega \eta \nabla u^f(x, \omega) + \nabla p_f(u) = 0,
\]

where \(\omega\) is the angular frequency, \(\eta\) the fluid viscosity and \(\kappa\) the frame permeability.

The following procedure was presented and validated in [8]. For the 2D case, on a reference square \(\Omega = (0, L)^2\) with boundary \(\Gamma\) in the \((x_1, x_3)\)-plane. We stand for \(\Gamma = \Gamma^L \cup \Gamma^B \cup \Gamma^H \cup \Gamma^T\), where

\[
\Gamma^L = \{(x_1, x_3) \in \Gamma : x_1 = 0\}, \quad \Gamma^B = \{(x_1, x_3) \in \Gamma : x_1 = L\}, \quad \Gamma^H = \{(x_1, x_3) \in \Gamma : x_3 = 0\}, \quad \Gamma^T = \{(x_1, x_3) \in \Gamma : x_3 = L\}.
\]

The boundary conditions are:

- Calculation of \(p_{33}(\omega)\).

\[
\sigma(u) \nabla \cdot u = -\Delta P, \quad (x_1, x_3) \in \Gamma^T, \quad \sigma(u) \nabla \cdot \chi = 0, \quad (x_1, x_3) \in \Gamma, \quad u^s \cdot \nu = 0, \quad (x_1, x_3) \in \Gamma^L \cup \Gamma^B \cup \Gamma^H, \quad u^f \cdot \nu = 0, \quad (x_1, x_3) \in \Gamma.
\]
Figure 3: Magnitude of reflection and transmission coefficients of the compressional wave as a function of the frequency and the incidence angle for a thickness $h = 0.001$ cm. The labels indicate $T =$transmission coefficient and $R =$reflection coefficient, $pf =$fast P-wave and $ps =$slow P-wave. The normal incident wave is a fast P-wave.

where $\nu$ is the unit outer normal on $\Gamma$ and $\chi$ is a unit tangent on $\Gamma$ so that $\{ \nu, \chi \}$ is an orthonormal system on $\Gamma$.

Denote by $V$ the original volume of the sample and by $\Delta V(\omega)$ its (complex) oscillatory volume change. In the quasistatic case, $\Delta V(\omega) = \frac{-\Delta P_{p33}(\omega)}{p_{33}(\omega)}$. Then after computing the average $u_{s,T}^{3}(\omega)$ of the vertical displacements on $\Gamma_T$, we approximate $\Delta V(\omega) \approx Lu_{s,T}^{3}(\omega)$ which enable us to compute $p_{33}(\omega)$.

- Calculation of $p_{11}(\omega)$.

This coefficient can be determine solving an identical boundary value problem than for $p_{33}$ but for a $90^o$ rotated sample.

- Calculation of $p_{55}(\omega)$.

The shear stresses allow to calculate $p_{55}(\omega)$ using

$$-\sigma(\omega) \nu = g, \quad (x_1, x_3) \in \Gamma_T \cup \Gamma_L \cup \Gamma_R, \quad u^* = 0, \quad (x_1, x_3) \in \Gamma_B, \quad u^i \cdot \nu = 0, \quad (x_1, x_3) \in \Gamma,$$

where

$$g = \begin{cases} (0, \Delta G), & (x_1, x_3) \in \Gamma_L, \\ (0, -\Delta G), & (x_1, x_3) \in \Gamma_R, \\ (-\Delta G, 0), & (x_1, x_3) \in \Gamma_T. \end{cases}$$

The change in shape suffered by the sample is $\tan[\theta(\omega)] = \frac{\Delta G}{p_{55}(\omega)}$, where $\theta(\omega)$ is the angle between the original positions of the lateral boundaries and the location after applying the shear stresses. Since $\tan[\theta(\omega)] \approx u_{1,T}^{s}(\omega)/L$, where $u_{1,T}^{s}(\omega)$ is the average horizontal displacement at $\Gamma_T$, $p_{55}(\omega)$ can be determined.

- Calculation of $p_{66}(\omega)$. 
Then, we rotate the layered sample 90° and apply the shear test as indicated for $p_{55}(\omega)$ to determine $p_{66}(\omega)$ (shear waves traveling in the $(x_1,x_2)$-plane).

### Calculation of $p_{13}(\omega)$

$$\sigma(u) \cdot \nu = -\Delta P, \quad (x_1,x_3) \in \Gamma^B \cup \Gamma_T, \quad \sigma(u) \cdot \chi = 0, \quad (x_1,x_3) \in \Gamma, \quad u^t \cdot \nu = 0, \quad (x_1,x_3) \in \Gamma.$$  

In this experiment $\epsilon_{22} = \nabla \cdot u^f = 0$, so that $\tau_{11} = p_{11}\epsilon_{11} + p_{13}\epsilon_{33}$ and $\tau_{33} = p_{13}\epsilon_{11} + p_{33}\epsilon_{33}$, where $\epsilon_{11}, \epsilon_{33}$ are the strain components at the right lateral side and top side of the sample, respectively. Then, since in this experiment, $\tau_{11} = \tau_{33} = -\Delta P$, and $p_{13}(\omega) = (p_{11}\epsilon_{11} - p_{33}\epsilon_{33}) / (\epsilon_{11} - \epsilon_{33})$. As the considered medium behaves as a VTI medium, let $\tau_{ij}$ denotes the stress tensor of this equivalent medium at the macroscale.

### 3.1 Numerical Results

In order to illustrate the procedure we consider fluid-saturated porous material with high fracture intensity, 20 fractures per meter, and low fracture intensity, 2 fractures per meter. The fracture thickness is 1 cm. These samples are characterized by a porosity $\phi = 0.25$ in the background and $\phi = 0.5$ in the fractures. The grain density is $\rho_s = 2650$ kg/m$^3$, while $K_s = 37$ GPa and $\mu_s = 44$ GPa are the bulk and shear moduli, respectively. Using the Krief model [14] we obtain $K_m = 1.17$ GPa and $\mu = 1.4$ GPa for the dry bulk and shear modulus of the background and $K_m = 0.58$ GPa and $\mu = 0.68$ GPa for the fractures. Permeability is obtained as $\kappa = r_g^2 \phi^3 / (45 - 45\phi)^2$ [15], where $r_g = 20\,\mu m$ is the average radius of the grains. The discrete boundary value problems to determine the complex
stiffnesses $p_{ij}(\omega)$ at the macroscale are solved for a finite number of frequencies using a public domain sparse matrix solver package. 

Using relations given in [13] it is possible to determine the energy velocities and dissipation coefficients from the $p_{ij}(\omega)$ for VTI media with horizontal symmetry axis.

Figure 5 shows polar representation of the energy velocities, where P is the quasi-compressional wave, SV is the quasi-shear wave and SH is the pure cross-plane shear wave. The curves correspond to a frequency of 30 Hz. In these cases we consider brine-filled and gas-filled fractures with a high permeability of the fracture material of $2.47 \times 10^{-10}$ Darcy.

It can be observed the anisotropy induced by fractures and the influence of the fluids. The same polar representation for the case with low fracture intensity and low fracture permeability ($2.47 \times 10^{-16}$ Darcy) is presented in Figure 6.

For completeness, Figure 7 shows the quality factor of the quasi-compressional and quasi-shear waves at 50 Hz. In this case with low fracture intensity, low fracture permeability and brine-filled fractures it is seen that energy losses are much higher for angles between 30 and 60 degrees for q-SV waves. Notice that for qP waves dissipation anisotropy is much important for waves traveling in the direction incident normal to the fracture plane. These upscaling experiments performed at the mesoscale allow to find the anisotropic coefficients in the constitutive relations of the equivalent viscoelastic medium at the macroscale.

4 WAVE PROPAGATION AT THE MACROSCALE

We presents the results of the seismic modeling in the macroscale obtained using a parallelizable procedure based in a nonconforming FE space.
We solve the following boundary value problem at the macroscale (in the 2D domain $\Omega$ with boundary $\partial \Omega$):

$$-\omega^2 \rho u - \nabla \cdot \tau(u) = F, \ \Omega \quad ; \quad -\tau(u)\nu = i\omega D u, \ \partial \Omega, \ \text{(absorbing boundary condition, } D > 0)$$

where $u = (u_x, u_z)$ represents the displacement vector, $\rho$ denotes the average density and $\tau(u)$ is the stress-tensor of our equivalent viscoelastic material, defined in terms of the calculated $p'_{ij}$s in the previous section using the upscaling procedure. Instead of solving the global problem associated with the above model, we obtain the solution using an iterative parallelizable hybridized domain decomposition procedure employing a nonconforming FE space.

The very simple computational model consists of an anisotropic layer at 500 m depth. This medium with vertical symmetry axis is the same material with high fracture intensity and high fracture permeability where the fluid that saturates the fractures is gas. The upper layer is an isotropic medium with P- and S-wave velocities equal to 2232 m/s and 625 m/s, respectively, and a density of 2100 kg/m$^3$.

The mesh contains 300×300 square cells having side length 5 m. The source is a dilatational perturbation indicated with an asterisk in the snapshots shown in Figure 8, where the interface is indicated. Its central frequency is 30 Hz. The vertical line, labeled A, represents a well at 150 m from the source location. The first receiver is located at 250 m above the first interface and the distance between adjacent receivers is 20 m. The solution was computed for 160 frequencies in the range 0–80 Hz. The snapshot to the left of the figure represents the vertical component of the displacement vector at 413 ms. The labels P, Pp y qP indicate P-wave, reflected P-wave and transmitted P-wave, respectively. The seismogram recorded at the well is displayed to the right. This numerical experiment was repeated for a anisotropic layer with horizontal symmetry axis, with the results being
Figure 7: Polar representation of the quality factor for qP and qSV waves at 50 Hz.

Figure 8: Snapshot of the vertical displacement at 413 ms (left) and seismogram recorded at the well (right). Transversely isotropic media with horizontal symmetry axis.

shown in Figure 9. The curvature of the wavefronts are a measure of the degree of the anisotropy of the medium.

5 CONCLUSIONS

The reflection and transmission coefficients for a plane fast compressional wave impinging on a fracture separating two half spaces within a poroelastic medium were computed and analyzed. The fracture was modeled as a thin layer or as a boundary separating the two half spaces, and the reflection and transmission coefficients were computed for both models. The fluid pressure can be equilibrated on both sides of the fracture if the fracture has high permeability and the layer thickness is small. In this case it is observed a good agreement among the reflection and transmission coefficients.

In addition, we present results from numerical rock physics. We obtained the com-
plex and frequency-dependent elements of the stiffness matrix for an effective fractured medium (transversely isotropic medium). In particular we have shown that the presence of fractures induces strong velocity and attenuation anisotropy that can be observed at the macroscale.

Moreover, we compare snapshots and synthetic seismograms for transversely isotropic medium with symmetry axes of different orientations. The modeling methodology (rock physics and wave propagation) is based on a finite-element solution of the equations of motion in the space-frequency domain. Hence, the frequency dependence of the seismic properties is modeled exactly, without approximations used in the time domain (e.g., memory variables based on mechanical models).

REFERENCES


AN HARMONIC 1D-ELEMENT FOR NON LINEAR ANALYSIS OF AXISYMMETRIC STRUCTURES: THE CASE OF HOT ROLLING

DENIS BENASCIUTTI, FRANCESCO DE BONA, MIRcea GH. MUNTEANU

DIEGM, University of Udine
Via delle Scienze 208, 33100 Udine, Italy
denis.benasciutti@uniud.it debona@uniud.it mircea.munteanu@uniud.it

Key words: Harmonic Model, Non-linear Analysis, Hot Rolling.

Abstract. Work rolls in hot rolling mills are subjected to thermal and mechanical loads that could affect the surface integrity of the roll. To estimate the roll lifetime, the thermomechanical behaviour of the work roll was recently studied adopting a 2D plain strain model. Nevertheless, even this 2D approach requires an extremely long computational time, which allows only thermal loads to be considered in the mechanical analysis. Work rolls are axisymmetric structures subjected to cyclic non axisymmetric thermal and mechanical loads, thus, to further reduce the dimension of the problem, an original 1D semi-analytical thermal and mechanical finite element is developed. Transient thermal and mechanical non linear analysis of a work roll is performed according to the initial stress method. The localized mechanical loads, which represents interaction of the work roll with the strip material and with the back up roll, are also taken into account.

1 INTRODUCTION

In hot mill rolling, work rolls are subjected to cyclic thermal and mechanical loads. Thermal stresses are caused by a non-uniform temperature distribution in the roll, as it is heated by the hot strip material and cooled by water sprays. Mechanical stresses, instead, are produced by interaction of the roll with strip and back-up roll. Such cyclic thermo-mechanical stresses can cause thermal and mechanical fatigue (e.g. "firecracks", spalling), wear and other damaging phenomena (e.g. oxidation at high temperature), which can lead to gradual deterioration of roll surface and even to failure.

For a proper design and maintenance of rolls, cyclic stresses in service have to be correctly estimated. For this purpose, numerical models based on the finite element method (FEM) have been proposed in the literature. Such models are characterised by different levels of complexity, depending on problem dimensionality (2D/3D models) and type of thermomechanical analysis. Some authors [1,2,3], for example, adopt a 3D model including strip plastic deformation and roll-strip thermo-mechanical interaction. This modelling approach allows a very realistic representation of the rolling process (e.g. roll bending, thermal deformation, strip plastic deformation). Such models are generally rather complex and then require a high computational cost, which may become unsuitable for everyday industrial
applications.

Alternative 2D models have been devised [4, 5, 6, 7, 8, 9], with the aim to simplify the analysis and to reduce the overall computational cost. These 2D models have shown the capability to represent with sufficient accuracy the thermo-mechanical response of the roll, which can help to understand the work roll service behaviour. On the other hand, results by 3D models confirmed that a 2D modelling could actually be realistic as well (3D models showed, for example, that the temperature is rather uniform along work roll axis, except a sharp drop in the regions close to strip edges).

In the 2D model proposed in [8], the strip is not modelled explicitly and it is replaced by suitable rotating thermal loadings. This approach permits good accuracy to be achieved, with a strong reduction of the computational effort. Despite the proposed modelling simplification, however, the transient thermo-mechanical simulations performed in [8] showed that even this 2D model still needs too long computational times, especially in mechanical analysis, where a non-linear solution is necessary to account for localised plasticity. This non-linear solution, for example, imposed a maximum number of roll revolutions to be simulated in transient mechanical analysis.

The aim of this work is then to explore if an approach based on semi-analytical finite elements could allow a further reduction of the computational time required by 2D models. This reduction would allow the user to simulate much longer time transients in mechanical analysis (until the steady-state condition is fully developed in the roll) by also taking into account mechanical loads.

Semi-analytical methods, developed more than fifty years ago, have been used to study axisymmetric structures loaded non-axisymmetrically, by means of 2D models [10]. Since work rolls have an axisymmetric geometry loaded by a non-axisymmetric thermal and mechanical loads, it seems possible to further reduce the 2D model of the roll to a 1D model.

In this work a 1D semi-analytical FEM approach will be proposed to analyze the thermal and mechanical transient behaviour of the roll. As it is well know, the semi-analytical approach is based on a Fourier series expansion, which leads to a linear system of uncoupled equations, due to orthogonally of harmonics. In this work, the presence of material plasticity, localized on the roll surface, requires a non linear analysis. The proposed 1D semi-analytical model, combined with the initial stress method, allows an efficient non-linear solution to be achieved.

2 SEMI-ANALYTICAL APPROACH

2.1 Linear case

A structure is axisymmetric if, in a cylindrical coordinate system \( r, \theta, z \) (where \( z \) coincides with axisymmetry axis) geometry, material properties and boundary conditions do not depend on the azimuth angle \( \theta \). Two loading conditions are possible. The first case concerns loads following the axisymmetric and the “anti-axisymmetric” distribution; the problem is therefore plane and results do not depend on \( \theta \) but only on \( r \) and \( z \). The second case occurs when the structure is loaded asymmetrically and the so called variable separation method [11] has to be used. Loads are developed in Fourier series:
where \( n \) is the harmonic order.

The response of a structure to a harmonic load is also harmonic and it can be obtained by superposing the response of the structure to each component of load. Displacements can thus be expressed as:

\[
\begin{align*}
\{u\}(r, \theta, z) &= \{u_0\} + \sum_{n=1}^{N} \left[ \{u_n\}(r, \theta) \cdot \cos n \theta + \{u_n\}(r, \theta) \cdot \sin n \theta \right], \\
\{v\}(r, \theta, z) &= -\{v_0\} + \sum_{n=1}^{N} \left[ \{v_n\}(r, \theta) \cdot \sin n \theta - \{v_n\}(r, \theta) \cdot \cos n \theta \right], \\
\{w\}(r, \theta, z) &= -\{w_0\} + \sum_{n=1}^{N} \left[ \{w_n\}(r, \theta) \cdot \sin n \theta - \{w_n\}(r, \theta) \cdot \cos n \theta \right]
\end{align*}
\]  

(2)

Generally only a finite number \( N \) of Fourier harmonics are needed to represent the load. It follows that, if the original problem has dimension \( p \), it can be replaced by a limited series of \( N \) analyses with reduced dimension \( p-1 \). This variable separation procedure is generally combined with FEM and called semi-analytical method. Different semi-analytical finite elements were developed in the literature [12, 13, 14] to reduce a three-dimensional (3D) problem to a plane (2D) problem.

In this work a simple formulation is proposed with the aim of reducing a 2D model to a 1D model.

In the following, the case of a 2 node semi-analytical finite element (also called harmonic element) is proposed; the same methodology could easily be applied to the 3 node element.

\[ \begin{array}{c}
0 \rightarrow 1 \\
L \\
2 \\
\end{array} \]

**Figure 1**: The 2 node semi-analytical finite element

Fig. 1 shows the proposed 2 node element. In-plane node displacements are given by the following expression:

\[
\begin{align*}
\{u\} &= \begin{bmatrix} u_1 \\ v_1 \\ u_2 \\ v_2 \end{bmatrix} = \sum_{n=1}^{N} \begin{bmatrix} \{\pi_{u,n}\} \cdot \cos n \theta \\ \{\pi_{v,n}\} \cdot \sin n \theta \\ \{\pi_{u,n}\} \cdot \cos n \theta \\ \{\pi_{v,n}\} \cdot \sin n \theta \end{bmatrix} + \begin{bmatrix} \{\pi_{2u,n}\} \cdot \cos n \theta \\ -\{\pi_{2v,n}\} \cdot \cos n \theta \\ \{\pi_{2u,n}\} \cdot \sin n \theta \\ -\{\pi_{2v,n}\} \cdot \sin n \theta \end{bmatrix} = \sum_{n=0}^{N} \left[ \{r_{1,n}\} \cdot \{\pi_{u,n}\} + \{r_{2,n}\} \cdot \{\pi_{v,n}\} \right]
\end{align*}
\]  

(3)

where:
\begin{equation}
[T_1] = \begin{bmatrix}
\cos n\theta & 0 & 0 & 0 \\
0 & \sin n\theta & 0 & 0 \\
0 & 0 & \cos n\theta & 0 \\
0 & 0 & 0 & \sin n\theta
\end{bmatrix}
\quad [T_2] = \begin{bmatrix}
\sin n\theta & 0 & 0 & 0 \\
0 & -\cos n\theta & 0 & 0 \\
0 & 0 & \sin n\theta & 0 \\
0 & 0 & 0 & -\cos n\theta
\end{bmatrix}
\end{equation}

In Eq. (3) simple and double over bars refer, respectively, to the symmetric and the anti-symmetric harmonic terms. A negative sign is introduced in front of the anti-symmetric terms as suggested in [14] only to formally simplify the equations that follow.

It can be noticed that the first harmonic, i.e. \( n=0 \), refers to both the axisymmetry and the "anti-axisymmetric" case. As only \( N \) harmonics are retained, the summation has to be limited to \( N-1 \).

The displacement of a point of coordinates \( r \) and \( \theta \) can be expressed as:

\begin{equation}
\begin{bmatrix}
u(r,\theta) \\
v(r,\theta)
\end{bmatrix} = [N]\{u^{el}\}
\end{equation}

where shape functions are simple linear relations:

\begin{equation}
[N(r)] = [N_1 \quad N_2] = \begin{bmatrix}
\frac{r_2}{r_2-r_i} - \frac{1}{r_2-r_i} r & -\frac{r_i}{r_2-r_i} + \frac{1}{r_2-r_i} r
\end{bmatrix}
\end{equation}

It is therefore possible to obtain the strain expression:

\begin{equation}
\begin{bmatrix}
\gamma_{rr} \\
\gamma_{r\theta} \\
\gamma_{\theta r} \\
\gamma_{\theta\theta}
\end{bmatrix} = \sum_{n=0}^{N-1} \begin{bmatrix}
[B_0] [\tilde{B}_1] [\tilde{B}_2] ... [\tilde{B}_{N-1}]
\end{bmatrix} \begin{bmatrix}
u^{el}_0 \\
\tilde{u}^{el}_1 \\
\tilde{u}^{el}_2 \\
\vdots \\
\tilde{u}^{el}_{N-1}
\end{bmatrix}
\end{equation}

where:

\begin{equation}
[B_0] = \begin{bmatrix}
\cos n\theta & 0 & 0 & 0 \\
0 & \cos n\theta & 0 & 0 \\
0 & 0 & \sin n\theta & 0 \\
0 & 0 & 0 & -\cos n\theta
\end{bmatrix}
\end{equation}

\begin{equation}
[B_n] = \begin{bmatrix}
\frac{\partial N_1}{\partial r} & 0 & \frac{\partial N_2}{\partial r} & 0 \\
-\frac{n N_1}{r} & N_2 & -\frac{n N_2}{r} & N_1 \\
\frac{n N_1}{r} & \left( \frac{\partial N_1}{\partial r} - \frac{N_1}{r} \right) & -\frac{n N_2}{r} & \left( \frac{\partial N_2}{\partial r} - \frac{N_2}{r} \right)
\end{bmatrix}
\end{equation}

and:

\( r_m = \frac{r_i + r_2}{2} \), \( L = r_2 - r_1 \).

Introducing the generalized Hook matrix:
the element stiffness matrix can be expressed as:

$$
[k] = \int_{-\pi}^{\pi} \left[ \begin{array}{cccc}
k_0 & 0 & 0 & 0 \\
0 & k_1 & 0 & 0 \\
0 & 0 & \overline{k}_1 & 0 \\
0 & 0 & 0 & \overline{k}_2
\end{array} \right] d\theta = \\
\left[ \begin{array}{cccc}
k_0 & 0 & 0 & 0 \\
0 & k_1 & 0 & 0 \\
0 & 0 & \overline{k}_1 & 0 \\
0 & 0 & 0 & \overline{k}_2
\end{array} \right]
$$

which is a $4(2N-1) \times 4(2N-1)$ matrix. Due to harmonic term orthogonality the stiffness matrix shows a semi-bandwidth of dimension 3. The standard FEM assembly procedure is finally used to obtain the structure stiffness matrix, which is still characterized by a semi-bandwidth of dimension 3.

In the case of the thermal problem a similar approach can be followed; obviously in this case each node is characterized only by a single degree of freedom. In particular in [15, 16] two approaches are proposed: a two-nodes finite element (with 1 or 2 Gauss point) and a three node element (with 2 or 4 Gauss points and quadratic shape functions). For the two-node finite element the “stiffness matrix” shows a semi-bandwidth of dimension 1.

Transient analysis is performed using the linear speed method (LSM) to integrate in time; as described in [16] this approach is unconditionally stable.

2.2 Non-linear case

In hot rolling, due to the high temperatures, a non linear analysis is required to deal with the elastoplastic behaviour of the material. As previously shown, in the linear case harmonics are uncoupled, due to orthogonality, and a narrow band stiffness matrix is obtained; such a favourable circumstance, however, does not occur in the elastoplastic case. It follows that, according to the standard Newton Raphson algorithm, the tangent stiffness matrix (Jacobian) has to be computed at each iteration for each harmonic, thus drastically reducing the efficiency of the computational procedure. A possible alternative approach, based on the initial stress method, was proposed in [17], for a non linear solution with a plane harmonic element. A similar procedure is proposed in [18] for a bilinear stress-strain relationship. Also in [19], the initial stress method is used in conjunction with a hybrid-stress functional based approach, to solve non-linear plane finite element models.

As a non-linear formulation for the one-dimensional case is not available, in the following the solution of the elastoplastic case according to the initial stress approach is described for the proposed 2 node harmonic element.

Considering a load increment, it is possible to write the generalized Hook law in term of...
stress and strain increments:

\[ \{\Delta \sigma\} = [D_{ep}]\{\Delta \varepsilon\} \]  

(10)

The expression of \([D_{ep}]\) is given in [20]:

\[
[D_{ep}] = [D] - \frac{[D]}{\sigma_0^*} \left[ \frac{\partial F}{\partial \xi} \right] \left[ \frac{\partial F}{\partial \xi} \right]^T \left[ \frac{\partial F}{\partial \sigma} \right] - 4 \frac{\partial F}{\partial \kappa} \frac{\sigma_0}{\sigma_0^*} + \left[ \frac{\partial F}{\partial \sigma} \right] \left[ \frac{\partial F}{\partial \sigma} \right]^T 
\]

(11)

where \(\kappa\) is the hardening parameter, corresponding to the cumulative equivalent plastic strain, which can be expressed as:

\[
\kappa = \varepsilon_{ep} = \int d\varepsilon_{pl} \approx \sum \Delta \varepsilon_{pl} \quad d\kappa = d\varepsilon_{pl} = \frac{2}{3} \left( (d\varepsilon_{pl}^e)^2 + (d\varepsilon_{pl}^p)^2 + (d\gamma_{pl}^{eg})^2 + 2\sigma_0^2 (\kappa) \right). 
\]

(12)

Symbol \(\sigma_0^*\) is the actual yield strength, while \(F\) is the yield criterion, which according to von Mises becomes:

\[
F(\sigma, \kappa) = (\sigma_x - \sigma_y)^2 + (\sigma_y - \sigma_z)^2 + (\sigma_z - \sigma_x)^2 + 6\tau_{xy}^2 - 2\sigma_0^2 (\kappa) = 0 
\]

(13)

It follows that, due to the expression defined by Eq. (10) of the Hook matrix, harmonics are no longer uncoupled in those elements where plasticity occurs. Consequently the stiffness matrix is not a band matrix but it has the following structure:

\[
[k] = \int_{A_i} \int_0^{2\pi} [D]\{\hat{B}\} d\theta \, dA + \int_{A_i} \int_0^{2\pi} \left[ D_{ep} - D\{\hat{B}\}\hat{B}\right] d\theta \, dA = [k_i] + [k_{nl}] 
\]

(14)

where:

\[
[k_i] = \begin{bmatrix}
    k_{0000} & 0 & 0 & 0 & 0 \\
    0 & k_{i000} & 0 & 0 & 0 \\
    0 & 0 & k_{i000} & 0 & 0 \\
    0 & 0 & 0 & k_{i000} & 0 \\
    0 & 0 & 0 & 0 & k_{i000}
\end{bmatrix}
\]

\[
[k_{nl}] = \begin{bmatrix}
    k_{11} & k_{12} & k_{13} & k_{14} & k_{15} \\
    k_{22} & k_{23} & k_{24} & k_{25} & k_{26} \\
    k_{33} & k_{34} & k_{35} & k_{36} & k_{37} \\
    k_{44} & k_{45} & k_{46} & k_{47} & k_{48} \\
    k_{55} & k_{56} & k_{57} & k_{58} & k_{59}
\end{bmatrix}
\]

\[
k_i = \int_{A_i} \int_0^{2\pi} \left[ D_{ep} - D\{\hat{B}\}\hat{B}\right] d\theta \, dA 
\]

Obviously, after the assembly process, the stiffness matrix of the structure \([K] = [K_i] + [K_{nl}]\) will show a similar characteristic.

According to the initial stress method, an iterative solution is performed by using the expression of the initial stiffness matrix, which is not upgraded in the following iterations. This approach can be profitably adopted also in the case of the 2 node harmonic element presented in this work, obtaining a system of uncoupled equations. It must be pointed out that, with respect to the non harmonic case, in which the bandwidth may have rather higher values, in the harmonic case an extremely narrow band matrix is always obtained, which obviously strongly reduces the computational time.
The computational procedure follows the usual iterative scheme. At \( q \)-th load increment the system of equations \([K_q]\{\Delta u_q\} = \{\Delta R_q\}\) is solved. \(\{\Delta R_q\}\) is the vector of unbalanced forces for the current iteration, which at the first iteration is equal to the load increment \(\{\Delta F\}\). The vector \(\{\Delta u_q\}\) contains the nodal displacement increments for all \( N \) harmonics. Strain increments, as well as the corresponding stress increments from Eq. (10), are then computed in the \( r \times m \) points of the domain (\( r \) nodes along the radius and \( m \) intervals in hoop direction). Stresses and displacements are actualized after each iteration: \(\{\sigma_q\} = \{\sigma_q\} + \{\Delta \sigma_q\}\) and \(\{u_q\} = \{u_q\} + \{\Delta u_q\}\). Inside each iteration a stress correction must then be applied only in the points where the Von Mises stress moves outside of the yield surface. For this purpose the value of \(\Delta \lambda = F_q \left( \frac{\partial F_q}{\partial \{\sigma\}} \right)^T [D] \frac{\partial F_q}{\partial \{\sigma\}}\) is computed, where \(F_q = F(\{\sigma_q\})\) is the Von Mises yield criterion. Consequently stresses are corrected iteratively \(\{\sigma_q\} = \{\sigma_q\} - \Delta \lambda [D] \frac{\partial F}{\partial \{\sigma\}}\).

At the end of the iteration the load vector equilibrating the actual stresses \(\{\sigma_q\}\) is computed: \(\{R_q\} = \sum \left( \int_{R_0}^{R_{\infty}} B_n \left[ B_1 \right] \left[ B_2 \right] \left[ B_3 \right] \ldots \left[ B_N \right] \right) \{\sigma_q\} d\theta dA\) and the unbalanced load \(\{\Delta R_q\}\) is then obtained. The next iteration of load step \(q\) is performed and so on until the load error \(\sqrt{\{\Delta R_q\}^T [D] \{\Delta R_q\}}\) is less than an adopted tolerance, when the next load step \(q+1\) is carried out.

3 THE CASE OF A WORK ROLL

3.1 Numerical model

The proposed semi-analytical model was applied to study the thermal and mechanical transient behaviour of work rolls in hot rolling mills. A work roll can be in fact considered an axisymmetric structure loaded asymmetrically. Geometry, loads and boundary conditions are constant along the roll axis, thus suggesting a 1D harmonic model to be used. In this work the results obtained with the proposed approach for the thermo-mechanical transient analysis will be compared with those obtained in [15, 21] with a plane model. Subsequently the effect of the mechanical loads due to the contact forces with the strip material and the back-up roll will be considered.

The thermal loads are represented in Fig. 1a. A cylinder of infinite length rotates uniformly and it is heated by a thermal flux that represents the contact with the hot material strip. A water spry cooling system is modelled by imposing convection on the arc defined by angle \(\psi\). The main simulation parameters are reported in Tab.1 and for sake of simplicity considered constant with temperature.

In the numerical analysis a thermal transient of 3600 seconds is considered, corresponding to about 1700 revolutions. The roll is supposed to be initially at room temperature. After the thermal analysis is performed, a subsequent mechanical solution allows obtaining stresses and
strains due to the temperature distribution already computed.

Table 1: Simulation parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R=300$ mm</td>
<td>Roll radius</td>
<td>$q_0=13.7 \times 10^6$ W/m$^2$</td>
<td>Thermal flux</td>
</tr>
<tr>
<td>$\omega=2.953$ rad/sec</td>
<td>Angular speed of the roll</td>
<td>$h=10100$ W/m$^2$K</td>
<td>Convection thermal coefficient</td>
</tr>
<tr>
<td>$\phi=10^\circ$</td>
<td>Heated sector</td>
<td>$T_0=20,^\circ$C</td>
<td>Water temperature</td>
</tr>
<tr>
<td>$\alpha=45^\circ$</td>
<td>Angular leg between hot and cold portions</td>
<td>$T_{roll}=20,^\circ$C</td>
<td>Initial roll temperature</td>
</tr>
<tr>
<td>$\psi=90^\circ$</td>
<td>Cooled sector</td>
<td>$\theta_p=6^\circ$</td>
<td>Angular distribution of contact pressure</td>
</tr>
<tr>
<td>$F=31400$ N/mm</td>
<td>Rolling force</td>
<td>$\theta_f=6^\circ$</td>
<td>Angular distribution of contact pressure</td>
</tr>
</tbody>
</table>

Fig. 2b shows the 2D finite element model (6940 elements, 6921 nodes) used for both the thermal and the mechanical analysis. Transient analysis is performed considering the roll fixed and the load rotating. A 4 node linear element is used for the thermal analysis, solution is obtained by means of an implicit algorithm based on a Jacoby gradient conjugate method. 609000 load steps were required, corresponding to almost 3 days of computation in a standard laptop. A 4 node isoparametric element in plane strain was used for the mechanical analysis. A kinematic material hardening model was considered. Solution was obtained by means of an explicit solver, with a modified Newton-Raphson approach. With respect to thermal analysis, computational time were much higher; only 20 roll revolutions were considered, with a computational time of 10 days. Further details are available in [8, 9].

Fig. 2c shows the model (57 nodes) in the case the 1D harmonic approach; the model is characterized by a node distribution along the radius, as that of the plane model; similarly harmonics function are evaluated at azimuth angles fitting the $m$ nodes distribution in hoop direction ($m=360$). As this 2 node semi-analytical element is not available in commercial codes, a self made finite element procedure was built up by using the Matlab programming language.

![Figure 2: Roll scheme (from [22]) (a); plane (b) and harmonic model (c)](image)

In an harmonic model, the number $N$ of harmonics that is retained in Fourier series expansion has to be chosen very carefully. As the thermal load is distributed over a very narrow angular sector ($\phi = 10^\circ$), a quite large number of harmonics was retained (50-80); it is therefore possible to achieve a good accuracy, avoiding errors related to Gibbs phenomenon,
with a quite reasonable computational time. In fact, with a standard laptop thermal analysis requires 1 h and the subsequent mechanical analysis 3 h.

If the mechanical loads have also to be considered in the model, an efficient solution procedure can be arranged by Fourier series expansion of an assumed pressure distribution. The contact region with hot material strip generally affects a quite narrow angle (e.g. for high rolling speed, it can reach values of 5°-7° [23]); to avoid Gibbs phenomenon, it is thus necessary to consider a great number of harmonics (N=80-100). In general, if a high number of Fourier terms have to be retained, the semi-analytical method reduces its computational efficiency. To overcome this problem, in the case of 2D harmonic finite element, in [4] the discrete Fourier series method is proposed. Nevertheless, if the 1D approach here proposed is used, the narrow band of the stiffness matrix permits high accuracy with still limited computational effort to be achieved.

3.2 Thermal analysis

Comparison between the solution obtained with the 1D and the 2D models was already performed in [16]. Only some relevant results are reported, with the purpose of better understanding the following mechanical simulation. Fig. 3a shows temperature distribution on the roll after 1800 seconds; along the roll surface, temperature increases reaching a maximum value of 388 °C in the hot zone.

![Figure 3](image)

**Figure 3**: Roll temperature distribution after 1800 s (a) and temperature versus time for points at different radial position (b)

Fig. 3b shows temperature variation versus time at different distance from the surface. It can be noticed that thermal gradients are localized in a limited portion of external surface. Plane and 1D harmonic models give similar results; obviously the computational time is strongly reduced in the second case.

3.3 Mechanical analysis:

**Thermal loads:**

Thermal stresses produced by the non-uniform temperature distribution calculated by thermal analysis are first considered. Figures 4a and 4b show the Von Mises stress after 20
roll revolutions (43 s), while Fig. 4c presents the stress variation in a point on roll surface during the first rotation. As already discussed in [9], the very high computational time required by mechanical analysis prevents the plane model to reach the complete development of the steady state condition in the roll.

Figure 4: Von Mises stresses (overall distribution and close-up view on the surface close to the thermal gradient): plane model (a), harmonic model (b) and Von Mises stress versus time in a point on roll surface (c)

With the purpose of comparing the two approaches, the same time interval was considered also for the 1D harmonic model, although a larger number of roll revolutions could be analyzed in this case. A very good agreement can be observed between the results obtained by the two methods. In particular, as already noticed in [16], comparing the stress distribution with temperature pattern shown in Fig. 3, it is possible to observe that stresses are mainly due to a constrained biaxial thermal expansion [24] and therefore a linear relationship occurs between temperature and stress.

Mechanical loads

Due to the interaction with the strip material and the back-up roll, work roll undergoes not only thermal, but also mechanical loads. According to [25, 5, 7], thermal stresses can be comparable or even higher than mechanical stresses. In [8,9], however, it has been shown that 2D model loaded only thermally is able to capture the relevant thermo-mechanical behaviour of the roll. The proposed approach, instead, allows mechanical loads to be also taken into account.

Tab. 1 lists the typical values of rolling force $F$ and angular sector $\theta_F$ for the contact pressure. In the bite region, the pressure is not constant and, in the following analysis, an elliptic pressure distribution is adopted. The pressure acts in radial direction, as friction is neglected. Pressure distribution $p(\theta)$ used in simulation has the following form:

$$ p(\theta) = \begin{cases} \tilde{p}(\theta), & \theta = \theta_0, \theta_0 + \theta_F \\ 0, & \theta \in [0, 2\pi) - [\theta_0, \theta_0 + \theta_F] \end{cases} $$

where $\tilde{p}(\theta)$ is the elliptical variation of contact pressure with azimuth angle. This pressure distribution is then expanded in Fourier series according to Eq. (1).

The rotating force is applied incrementally; at each time step $\Delta t$ the angle $\theta_0$ (see Eq. (15)) is increased of $\Delta \theta_0 = \omega \Delta t$ where $\omega$ is the angular speed of the rotating force. The Fourier expansion of the rotating force is performed at each time step.
Fig. 5a shows the Von Mises stress at $r/R=0.95$ versus time within a complete revolution. Figure 5b shows the Von Mises stress due to the rotating mechanical load after a revolution of $\pi/2$. It has to be noticed that in this case thermal loads are not present; due to the mechanical load, localized plasticity occurs and, therefore, residual stresses are present also when the loads drift apart. Fig. 5c shows an example of result obtained considering both thermal and mechanical loads. While the maximum thermal stress occurs on the roll surface, the maximum mechanical stress, instead, is localized in a inner portion of the roll, beneath the surface.

4 CONCLUSIONS
- In this work, a 1D thermal and mechanical harmonic element is proposed for analyzing structures that can be reduced to a plane axisymmetric model with non-axisymmetric loads.
- The proposed element is suitable to describe the behaviour of a work roll in hot rolling; in fact, in order to evaluate the plastic behaviour of the component, a non-linear transient thermo-mechanical analysis has to be performed, which requires a high computational effort also if a 2D model is considered and mechanical loads are neglected. With respect to the 2D approach, the proposed 1D model permits a strong reduction of the computational time to be achieved.
- According to the proposed procedure, the effect of the mechanical loads can also be easily taken into account.

REFERENCES


ON USING PETRI NETS TO MODEL AND SIMULATE PROCESS WITH UNCERTAINTY

César Monteiro Ribeiro*, Adriano Chaves Lisboa†, Carlos Andrey Maia*, Rodney Rezende Saldanha*

*Graduate Program in Electrical Engineering - Federal University of Minas Gerais
Av. Antônio Carlos 6627, 31270-901, Belo Horizonte, MG, Brazil
cesar@cpdee.ufmg.br
maia@cpdee.ufmg.br
rodney@cpdee.ufmg.br
www.cpdee.ufmg.br

†ENACOM - Engenharia Assistida por Computador, Brazil
770 Professor José Vieira de Mendonça St., Belo Horizonte MG 31310260
adriano.lisboa@enacom.com.br
www.enacom.com.br

Key words: simulation, stochastic Petri nets.

Abstract. This work presents a study on using Petri nets to model and simulate processes with uncertainties. They typically describe and simulate discrete event systems, especially queue analysis which is of great interest in processes. Their measures are useful for understanding the process and they naturally compose formulations of optimization problems for a process. Petri nets are notable for their simplicity and elegant mathematical support. Their simulation involves a few simple matrix equations and graph operations, so that their analysis has been explored in depth. The Petri net fundamental concept is quite simple: tokens that go from one place to another place as transitions connecting those places fires after having enough tokens for that. This work instantiates and analyses a simple process with uncertainty in order to properly show the applicability and usability of Petri nets using a structural approach.

1. INTRODUCTION

The process of running a computer model of a system defines simulation. It allows predicting the behavior of a modelled system under given conditions using a computer [1]. A simulation imitates, or simulates, the real world systems or processes. A system can be defined as an aggregation of objects that have some interaction and interdependence. These systems have in common the way they realize occurrences around them, which is given through events. These events are instantaneous, which gives them a discrete feature in time. This kind of system is called discrete event system (DES), as opposed to continuous variable systems, treated by classical control theory. The discrete nature of DES's makes conventional mathematical models based on differential equations unable to solve them.

A discrete event system can also be dynamic, i.e. it changes with the occurrence of events, which usually happens in unknown irregular time intervals. The occurrence of an event causes
a state transition [1]. SED may be applied to manufacturing, staffing personal-service operations, post office, distribution and logistics, health care, computer system and telecommunications, military and public policies.

This work investigates Petri nets. A Petri net is an elegant way to model a SED and it allows a fast simulation. We demonstrate its capabilities by example, applying it for modeling an open pit mine.

2. SIMULATION WITH PETRI NETS

The theory of Petri nets was introduced by the German Carl Adam Petri in his doctoral thesis entitled “Communication with Automata” in 1962 at the Faculty of Mathematics and Physics University of Darmstadt.

A Petri net is a system specification technique that allows a mathematical representation and analysis that has the ability to verify properties and is able to check consistency of a specified system. With the addition of time concept, it is possible to represent the dynamic behavior of systems that have concurrent, asynchronous, distributed, parallel, non deterministic, and/or stochastic [3] state changes. Petri nets, as graphical and mathematical tools, provide a uniform environment for modeling, formal analysis, and design of discrete event systems. Furthermore, they allow visualizing processes and communication between them.

Petri nets are formed by two types of components: one active and one passive. The active is called transition and the passive is called place. Places represent the state variables and transitions represent to actions performed by the system. Figure 1 shows places represented by circles and transitions represented by lines or bars.

![Figure 1: Places (circles) and transitions (bar) of Petri nets.](image)

An action is associated with some conditions (state variable), i.e. there exists a relation between places and transitions that enables an action. Similarly, after an action, some places change their information.

When a transition is enabled, it can fire, i.e. an action can happen. A transition becomes enabled when the number of tokens at input places is larger than a given threshold, specified by arc weights. Firing a transition consists of two operations: removing from each input place the number of tokens given by the weight of the respective arc, and adding to each output place the number of tokens given by the weight of the respective arc. Firing of transitions goes on while there is at least one enabled transition. If there is no enabled transition, the execution will be finished. An action in the system (firing of a transition) is associated with
model-specific features and it mathematically follows a probability distribution that best approximates the behavior of this action.

Petri nets have been applied to a transport system [6], where they modeled traffic lights in urban centers. In this case, the model was built based on the concept of structural approach [7], which uses Petri subnets in the construction of the final model so that changes in one of the subnets do not affect others. Another approach of the structural concept for Petri nets is applied to the Ethernet protocol [8]. In this scenario, it is possible to manage the protocols and services in a structural way, piecewise, without interference of adjacent services. Petri nets have also been applied to software engineering processes [9].

3 Modeling with Petri Nets

After describing how to simulate systems with Petri nets, the next step is to describe how to model a system with places and transitions. Inputs of a transition are the preconditions of the corresponding event. The outputs are post-conditions. An event corresponds to firing the respective transition. A true condition is represented by tokens in input places that form the condition. When a transition fires, it removes the tokens that represent the real preconditions and creates new tokens that represent the transformation of the corresponding true post-conditions [3].

Scheduling of firing times follows probability distributions associated with the nature of the events. After each update of tokens along the network, just enabled transitions schedule their respective firing in a schedule event list, where all scheduled events are stored together with their respective firing time.

This paper presents an application where a stochastic timed Petri net model and simulates system where resource sharing and parallelism are present. In these cases, a Petri Net is a simple and efficient tool when compared to other tools of discrete event systems because it allows visualization of the process dynamics and it provides respective measures of performance.

3.1 Case study

As a case study, we apply Petri nets to model and simulate open pit mines in order to obtain its productivity. The model consider typical loading and hauling processes for open pit mining operations.

The block diagram in Figure 2 represents a mine with loading and hauling process.
Figure 2: Loading and hauling processes.

This system was modeled and simulated by the Petri net shown in figure 3. The places represent the pre and post conditions of the resources: trucks, loader, loadout. The transitions represent the occurrence of events during the simulation process.

Input queue: initial condition of the model, where truck attributes are defined (e.g. transport capacity, average operating speed, fuel consumption). A token in the input queue place indicates that a truck is ready to start or restart new loading and hauling process.

Submodel load square: the trucks are positioned on the load square to start the loading process. The transition “start moving load square” is enabled if there is a truck in the input queue. This transition fires after a delay associated with the time spent by the truck in the load square. At the end, one token will be add in the next place: “truck ready to load”.

Submodel loader: this submodel simulates the truck loading. The transition “start loading” may be enabled and will be fired immediately if the following conditions are met: there is a truck ready to be loaded and the loader is idle. After the transition is fired, two post-conditions are updated: “busy loader” receives a token and “idle loader” looses a token, disabling the transition “start loading”. This is a way to limit the service to only one truck at a time. The “busy loader” place enables the transition “finish loading”. This transition simulates the loading time. The loading time has a specific probability distribution that models the loader behavior to reach the minimum of 90% of the maximum load capacity of the truck. After firing “finish loading”, the loader goes back to idle state and the truck is loaded, as indicated by the respective places.
Submodel loadout queue: this block is analogous to the submodel load square queue, where the truck is positioned to discharge.

Submodel loadout: this submodel simulates the truck discharge. The spent time in discharge also follows a specific probability distribution that models the behavior of the loadout to receive the load of the truck. The transitions firing process follows the same logic as the loader submodel, which is limited to only one truck at a time.

Transition “send truck to input queue”: it is enabled when there is an unloaded truck. The delay in this case corresponds to the truck travel time between the hauling and loading processes.
Table 1: Probability distribution for processes.

<table>
<thead>
<tr>
<th>Process or equipment</th>
<th>Parameter</th>
<th>Probability distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Load Square</td>
<td>Moving time</td>
<td>Triangular</td>
</tr>
<tr>
<td>Load</td>
<td>Loading time</td>
<td>Triangular</td>
</tr>
<tr>
<td>Loader</td>
<td>Load</td>
<td>Triangular</td>
</tr>
<tr>
<td>Moving full</td>
<td>Travel speed</td>
<td>Normal</td>
</tr>
<tr>
<td>Outload queue</td>
<td>Moving time</td>
<td>Triangular</td>
</tr>
<tr>
<td>Outload</td>
<td>Discharging time</td>
<td>Triangular</td>
</tr>
<tr>
<td>Moving empty</td>
<td>Travel speed</td>
<td>Normal</td>
</tr>
</tbody>
</table>

4 SIMULATIONS RESULTS

The parameters used in the simulation are listed in the table below:

Table 2: Parameters of the probability distribution for the process

<table>
<thead>
<tr>
<th>Process or equipment</th>
<th>Parameter</th>
<th>Probability distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Type</td>
</tr>
<tr>
<td>Load square</td>
<td>Moving time(min)</td>
<td>Triangular</td>
</tr>
<tr>
<td>Load</td>
<td>Loading time(min)</td>
<td>Triangular</td>
</tr>
<tr>
<td>Loader</td>
<td>Load(ton)</td>
<td>Triangular</td>
</tr>
<tr>
<td>Moving full</td>
<td>Speed(km/h) %</td>
<td>Normal</td>
</tr>
<tr>
<td>Loadout queue</td>
<td>Moving time(min)</td>
<td>Triangular</td>
</tr>
<tr>
<td>Loadout</td>
<td>Moving time(min)</td>
<td>Triangular</td>
</tr>
<tr>
<td>Moving empty</td>
<td>Speed(km/h) %</td>
<td>Normal</td>
</tr>
<tr>
<td>Mean Speed</td>
<td>Speed(km/h) %</td>
<td>Constant</td>
</tr>
<tr>
<td>Mean load</td>
<td>Load(ton)</td>
<td>Constant</td>
</tr>
</tbody>
</table>

The following are the results of the simulator and its comparison with the reference model of the Arena [5]. ARENA is a graphical computer simulation interface, which uses discrete event simulation.

The annual production represents a total moved of mass in toneladas in the simulation period. This is a sum of an individual production of each truck. Considering $M$ trucks, which and will perform $N$ complete cycles in the simulation period, and the variable $C$ is a random variable which represents the load of the truck in each cycle. The equation of production is:

$$Production(ton) = \sum_{i=1}^{M} \sum_{j=1}^{N} C_j$$  \hspace{1cm} (1)

The table 3 presents the simulation results; we observed a difference of 2.13% for the simulation with 1 truck. The simulation period was 1 year and 15 replications were performed. The simulation time parameters and number of replications were adopted in order to enable
comparison with the reference model of the Arena [5].

Table 3: Comparing Petri nets Simulator x Arena

<table>
<thead>
<tr>
<th>Production (mil toneladas)</th>
<th>Truck numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Arena</td>
<td>4.682</td>
</tr>
<tr>
<td>Petri Net Simulator</td>
<td>4.582</td>
</tr>
<tr>
<td>Difference %</td>
<td>2.13%</td>
</tr>
</tbody>
</table>

The graphs below show the differences between the models, showing the comparison for each one of the 15 replications. As each replication is equivalent to 1 year of simulation, then we have 15 years of simulation and their respective differences in production in tons.
5 CONCLUSIONS

During the development of this approach some simulation tools was found in the literature. They are applied in academics and commercial applications. But some of them are protected and did not allow parameters to be changed. So, we decide do develop a simulator that allows total control over the simulated processes and that has equal or better performance to those found in the literature. This paper provides us some contribution; one of them was the development of a tool: the simulator of Petri nets and other contribution we can highlight were the functions of transitions fire that were constructed to model the specific characteristics of the events of the model, showed lower variability in the value of production compared to Arena, as well as higher processing speed of the results. We observed that the operating process studied has a number of variables with stochastic features, such as production processes involving people. For future studies will be considered other objective of the production environment what is, to reduce the time that the equipment is in nonproductive activities.

6 REFERENCES


**ACKNOWLEDGMENTS**

CAPES  
CNPq  
Fapemig  
Graduate Program in Electrical Engineering - Federal University of Minas Gerais  
This work has been supported by the Brazilian agency CAPES
REVERSE ENGINEERING AND FORMING SIMULATION OF INDOOR OF AN AUSTENITIC DISHWASHER

JUAN C. CISNEROS¹, JOSÉ L. MACÍAS² EMILIO ROYO³

¹Universidad Popular Autónoma del Estado de Puebla, México
21 sur 1103 Barrio Santiago C.P. 72410, Puebla, México
juancarlos.cisneros@upaep.mx

² Universidad Popular Autónoma del Estado de Puebla, México
21 sur 1103 Barrio Santiago C.P. 72410, Puebla, México
joseluis.macias@upaep.edu.mx

³Universidad de Zaragoza, España
María de Luna 3, 50018 Zaragoza, España
eroyo@unizar.es

Key words: Austenitic Stainless Steel, Forming, FEM, Reverse Engineering.

Abstract. Majority of automotive, aerospace and appliances components have been designed for a methodology based on metal forming simulation. Besides this, there is a growing demand for new products, with more complex geometric design and production time shorter. Fortunately, nowadays we have more tools intended sheet metal forming simulation and reverse engineering like HyperWorks and Geomagic Design.

In this paper a real case of electrical appliance component “Door Lining of an austenitic dishwasher” is discussed. The main objective is to develop a methodology to combine and evaluate several tools employed in the Reverse Engineering and the Forming Simulation. The basic steps used for conducting the door lining forming simulation, including the reverse engineering process, are given next.

The process begins with the 3D scanning and point cloud processing to generate a CAD Surface; furthermore the cad model inspection and quick feasibility study for the blank development; besides that, the tools generations for the triple action draw to the incremental formability analysis; finally, the thickness comparison and whole deviation between the simulation results and the real part.

1 INTRODUCTION

In the last six years Mexico has been catapulted as a global manufacturing star due to the improvement in its competitiveness index and its low production costs. Hand in hand with this, the appliances industry represents a third of the foreign investment in the manufacturing
sector, while ten years ago it meant only the 8%. Making the parts for appliances industry means adding stamping operations and forming more complex figures.

In the developed world of manufacturing, cold stamping is an undisputable reference of forming processes due to the number of parameters that need to be controlled in its actions and its sensitivity for determining the success or failure of the process. In a similar way, challenges associated with high precision geometries, tolerances and high speed productions must be faced.

The requirement of owning production processes where the number of failures is the minimum is essential in the appliances sector, consequently, design and manufacturing engineers, adopt more and more, ultimate generation methods that let them solve different challenges in designing and manufacturing of metal sheet stamping. Meanwhile, geometries complexity, high precision grade and sheet’s properties, generate the need of calculating physic variables as effort, friction coefficients, displacements, looseness, among others, all related to the process itself; due to this, forming engineers must be capable of foreseeing the process’s conditions and variables [1].

Numeric simulation based on finite elements analysis is the most developed procedure for controlling and foreseeing those conditions and variables that affect stamping processes in a negative way. Their objective lies in observing, before prototypes generation or fabrication, metal sheet’s behavior during stamping process and determining its sensitiveness before diverse variables. Thence, in recent years the state of art of metal sheet’s transformation by deformation is related to the level of development of the technology employed in finite element simulation [1].

Although the tools of finite element simulation are powerful and reliable for designing and production engineers, it is unavoidable that conditions such as lubrication, engineer’s expertise and holding conditions that cannot be simulated exist and determine radically the results’ degree of reliability. However, diverse technologies that support numeric simulation had developed.

One of them is inverse engineering, this one allows doing a systematic approach of products and systems’ designing; in which one piece, previously manufactured, can be geometrically evaluated and be re-manufactured. The inroad of new instruments in three-dimensional digitalization has facilitated inverse engineering usage also, in its higher level; it is possible to perform a piece’s dimensional inspection in real time, calculating its deviations and deformations caused by process’s own designing. Due to that, inverse engineering use through three-dimensional digitalization and the results of numerical simulation by finite element analysis, can be syncretized with the goal of verifying the results and give more validity to the overall process [3, 4, 5, 6].
2 METHODOLOGY

2.1 Overview of Computer Simulation in Sheet Stamping Processes (CAE)

Nowadays many companies dedicated to metal sheet stamping are discovering that, even though working with a high efficiency grade in traditional techniques, it is not possible to grant its success in a market that demands very short deadlines or “times to market”, due to the existing competition in the market, utility margins depend more and more on who reaches first the customer [6].

By the above, companies have focused their efforts in the application of CAE (Computer Aided Engineering) tools in forming processes as a tool that streamlines designing problems solution and production processes’ optimization. Roughly, Computer Aided Engineering allows forming processes simulation by the using of powerful commercial software modules which are based in the principle of finite elements [2, 3, 6].

CAE’s value, in industrial sector, lies in its simplicity to quickly determine the viability in the manufacture or design of pieces and to redesign them. With it, it is possible to reduce the number of tests and correction cycles related to the classic methodology of “proof and error”, driving to a methodology of “correct since the beginning”. Likewise; CAE is capable of responding to the classic “what if...?” design questions, that allows designers to evaluate different options and explore alternatives during the designing stage [6].
CAE methodology tends to increase costs and the employed time for development of first prototypes, since ideas are tested virtually before being employed in the workshop. Even so, the increasing of designing time is compensated afterwards reducing the number of tests in prototypes and the problems in production. The time and cost for solving a problem increases 10 times depending on the stage of product’s development where it is identified; for example, if the problem is found in the concept’s definition stage, the cost and time will be 1, but if it is found on the designing stage it will be 10, on prototype’s fabrication stage 100, on prototype’s testing stage 1000, and on fabrication stage it will be 10000 [6].

As it has been said, CAE simulations allow the reduction of delivery times, problems’ anticipation and the taking of preventive decisions. In forming processes case, CAE/CAD technologies allow:

- The quickly testing and sketching of pieces,
- The specification of the design of the piece,
- The simulation of deformations to which the piece will come under,
- The definition of its geometry,
- The definition of the geometry of the dies for its construction.

Among other advantages, because CAE applications are very intuitive tools, an expert may not be necessary for their use. CAE tools have been used as correction tools; however, their competitive advantage is found in prediction stage [3, 6].

Figure 2: Sheet Metal Model Deforming Process Simulation by Means of CAE
2.2 Simulation Methods

The simulation methods employed in metal sheet forming processes are two: Inverse method or OneStep and incremental analysis method. OneStep analysis is used in the product’s developing stage and in the initial steps of the process’s development; its final goal lies in obtaining viability and conformability of a project. The second one is applied in simulation process for completely defining tools’ parameters and the forming process [1, 2, 6].

2.3 OneStep

The codes of OneStep are easier, faster and cheaper in the acquisition of solutions. Nonetheless, tooling is not contemplated in it, which makes impossible for the software to find answers to the progression of wrinkles generated in the forming stage. Because of it, this method is used only on one-step stamping processes.

For the realization of this simulation the following requirements are needed:

- CAD geometry or the mesh of the final piece
- The geometry of the finished piece with the tread zone
- The data of the material as thickness, Young’s module, Poisson coefficient and the Tension-deformation curve.

The main advantages of OneStep are:

- Early indication of the piece’s formability, before the forming process is defined
- Reduction of invested time in the development of prototypes and testing
- Approximate determination of the material’s thinning percentage
- The initial development of the process and its optimization ways, such as: brakes positioning or the establishment of treading forces
- Easy virtual modification of the product.

2.4 Incremental

The incremental analysis allows the simulation of multiple stages and the obtaining of multiple increases of each one of them. Thence, process’s evolution can be observed and the stage where the failure or excessive wrinkling can be determined [2, 6].

The requirements for the realization of the incremental analysis study are:

- CAD geometry or the mesh of the sheet’s initial format,
- The geometry or mesh of the tooling,
- To stablish parameters as, the number of steps, the tooling movements, treading forces, trims and restriction forces,
- Material data as thickness, Young’s module, and Poisson’s coefficient, directionality, friction coefficients and the Tension-deformation curve.

The advantages of the incremental method are:

- Reduction of times and costs of the product’s development,
- Early problems indication in the forming process,
- Quickly obtaining of more precise results related to the brakes’ effect, thinning, wrinkles and more,
- Optimization and validation of the process,
- Cracks and wrinkles’ determination,
- Obtaining of the FLD diagram,
- Determination of the elastic recovery of the piece to be formed.

2.5 Proposed methodology:

The method developed in this thesis consists of a sequence of operations limited by five phases that together establish the development of a dishwasher’s inner austenitic door simulation, as shown in the Figure 3 diagram [6].
The process begins with the three-dimensional digitalization of the dishwasher using an optic no-contact scanner HANDY SCAN 3D; from which some facets are obtained and later converted in a dense point cloud that represents the objective’s geometry in Vx Elements software, as shown in Figure 4, these points are exported to a specialized virtual modelling software CARE [4, 5, 6].

Figure 4: Phase 2. Virtual Modelling

In this phase the caption of the intensity and the designing parameters of the original piece are established, which lets the CAD model to be modified by the process’s requirements. Due to the piece’s geometry, it is necessary to use the Autosurfacing Module of Geomagic Dx because it adjusts NURBS surfaces to the point cloud obtained in the previous phase. Likewise; information referring to position and geometry is obtained from the drawbeads in this phase [4, 5, 6].

Figure 5: Phase 3. CAD model generation

In CAE simulation the type and number of employed tools in the stamping process is defined; in the first instance, with OneStep the original preform and the conformability of the piece are established; later, with Die process tooling for triple effect stamping process are generated; without forgetting, gravity’s effect and preform’s springback. As that being said the following configuration of multiple stages simulation in HyperForm is proposed [6].
In this phase the dishwasher’s inner door finished part after the cut is digitalized and is compared with the CAD geometry obtained from the multi-stage stamping process of phase 3. Originating a report of the existing deviations between the CAD model and the point cloud that will determine the reliability of the project [4, 5, 6].
3 Main Results

One of the goals for stamping simulation is to foresee defects in the forming process and, in the same way; it can help to validate the product’s design feasibility [6].

- The observed results from the One Step analysis correlate well with those from Incremental analysis.

- The blankholder friction, mesh size and punch velocity are the most influential factors in the thinning and formability of the part.

- Thickness showed a greater difference: the maximum thinning from the simulation was 21% while thinning from the real model measurements was 35%.

4 Conclusions

In this investigation paper’s development different activities have been described and proposed, based on and structured methodology, that allowed the dynamic and complete simulation of a domestic dishwasher inner door stamping process. The results obtained during this work generate a set of conclusions and suggestions that will be outlined below.

- Referring to the three-dimensional acquisition stage, a wide knowledge about the laser scanning equipment (Handyscan) and the controlling of its respective software (Vx Elements) was obtained, it can be observed of them that the digitalization and facets optimization process is very flexible and reliable for inverse engineering activities; even average digitalization times are lower compared to the typical systems used in instrumentation.

- In order to capture completely the geometry of digitalized models, many different activities were developed. From these, it can be noted that one of the factors that generate more noise in a three-dimensional optic or by laser scanning system, is the surface’s reflectance, causing a pretreatment based in the clouding of the surface by the superficial application of powders or controlling the luminosity of the place where digitalization is performed.

- To feed the CAE software with CAD surfaces, it was performed in Geomagic Dx the transformation from .STL files to native CAD formats .iges, through system’s inverse engineering. Consequently, CAD geometries were obtained from the digitalized model with a high precision level and some sections fully parametric, allowing in the future the partial redesigning of the original model.

- The value of working with software based on the finite elements analysis of manufacturing applications such as stamping, lies in the possibility of foreseeing the behavior of a metal sheet before some conditions and parameters before a product is
manufactured; hence it is a primary and added value part for validating the proposed
designing global process.
- When finite elements simulations are performed, the input parameters must be
idealized, and these depend on the engineer’s ability, knowledge and experience;
likewise it must be clear that simulation is just an approximation means to the good
results and that it cannot substitute in no way the lab tests. With that being said, the
more knowledge you have about the numeric limitations involved in the process,
security in the final results will be greater.
- For increasing the degree of confidence of simulation real material physical
parameters must be used, generally their characterization for effects of simulation is
based on the established properties by a supplier; however, care must be taken in their
reliability because it can generate unreal simulation results.
- Hyperperform demonstrated that it is a CAE software specialized in sheet deformation
processes and turned out being a useful tool for stamping processes’ simulation, since
it allows access in a simple and graphic way to the problems that may be presented,
and quickly determine the corrective actions by the readjustment of parameters.
Besides, it presents a global stamping stage such as Multi-Stage manager, where
forming analysis is performed including gravity and elastic recovery effects that allow
to simulate the whole stamping process in a more precisely way. Nevertheless, the
software presents very high calculating times for big geometries causing that the size
of the mesh and the proposed forming speed for fast runs are greater than the usual.
- In the Multi-stage manager numeric simulation development, it is essential to have a
broad knowledge about the Hyperperform working environment due to the fact that the
generated result in a stage is used as an input for the next one; besides knowing the
typical importing and exporting formats supported by the software.
- From the parameters adjustment performed by the stamping simulation; it can be
observed that meanwhile the size of the mesh tends to be smaller, the simulation
results converge in a better way; likewise, it was observed that the friction
coefficient’s variation does not affect its elastic recovery, this parameter must be
controlled cautiously as it may cause failures in the material; and finally, the incursion
of the brakes’ geometry in the tooling, is a simple way of reducing elastic recovery
and boost the thinning in zones where it is necessary to improve stamping quality.
- For evaluating MEF simulations performed in this project, two methods were
established for studying the behavior of the two most important variables in the paper.
The first one, thinning, that was evaluated through HyperView simulation and the
thicknesses physical acquisition. The second one, form; in this one form deviations
generated between ideal digital models can be observed, as a result of the simulation
after the trimming, with the final piece’s digital model obtained by the scanner. Of
these, two important conclusions were obtained, that the proposed simulation
methodology foresees thinning percentages lower that the obtained in the original production of the piece, and the second conclusion is that it has been proved that the geometric form of the piece has an elevated correlation with the final piece.

- In the inspection area, Geomagic Verify turned out to be an excellent tool for results validation processes, due to the fact that it counts with infinite tools for parts’ aligning, saving time in the pre-processing and giving reliable deviation results.

- Generally it was achieved to generate knowledge about the management of computer systems for engineering, since the generation of acquisition and complete simulation of stamping process was possible with the help of typical engineering design software, also the use of new technologies employed in inverse engineering. Likewise, it is demonstrated that with the proposed methodology excellent results were obtained, even though it is a simulation project, it can be used as a tool for reducing the testing phase and for getting shorter product’s development cycles.

- Finally, this paper provides important foundation to continue with the simulation of other sheet forming processes with inverse engineering and to be a starting point that opens new doors to the simulation of phenomena such as elastic recovery and the gravity effect in the stamping process.

REFERENCES


NON-IONIZING RADIATION MAPPING BASED ON GEOGRAPHICAL INFORMATION SYSTEMS

OSCAR VALENCIA, JUAN HUIZA, MAURICIO POHL AND ENMANUEL AMAYA

Departamento de Electrónica e Informática Universidad Centroamericana “José Simeón Cañas”
Bulevar de Los Próceres, San Salvador, El Salvador, Central America
ovalencia@uca.edu.sv, jhuiza@uca.edu.sv, mapohl@uca.edu.sv, earaujo@uca.edu.sv

Key words: DTM, TIN, Ray-Tracing, Electromagnetic wave propagation, Geometric optics, Diffraction theory.

Abstract. This document presents the results for an implementation of geo-referenced Digital Terrain Models (DTM), designed to show tridimensional maps with information related to the non-ionizing radiation intensity in the surrounding areas of one or more half wave antennas. A procedure that makes use of a software tool developed to apply a Ray-Tracing based algorithm for the calculation of wave propagation variables. This algorithm is handled as an asymptotic optical method specialized in high frequency fields.

1 INTRODUCTION

In El Salvador, the advance of the technological development and the rising demand for mobile telecommunication services have generated a need in mobile phone companies to increment their infrastructure in order to maintain a high quality level of service [1].

However, general population and local authorities have manifested a constant concern for the possible health issues the radiation originated from these infrastructures can cause, especially in rural areas where measuring its effects is more difficult [2].

In response, the Superintendence of Electricity and Telecommunications (SIGET), together with the International Union of Telecommunications (UIT) and the Regional Technical Commission of Telecommunications (COMTELCA), has implemented the Non-ionizing Radiations Project, which objective is to measure and monitor the electrical field intensity values all throughout the country, to control that these values do not exceed the norm established by the International Commission for Protection against Non-ionizing radiations (ICNIRP) [3], and to maintain the population informed about these results to increase their trust in the proper functioning of the telecommunication nodes.

As part of said project, the SIGET has already generated radiation charts for some municipalities, mainly inside the capital San Salvador. The measurements needed to create these charts were mostly taken by a specialized device attached to a car that travels across the relevant zones receiving electrical field intensity data at several points. However, most of the time, these measurements are limited only to main roads or small areas that the car can access, so electrical field values for other sectors cannot be approximated with precision. Another
difficulty arises from the fact that these charts need to be updated as much as possible, and with such equipment this process can be tedious.

The objective of the exposed implementation is to able to present, via digital mappings, accurate predictions for the non-ionizing radiation intensity and electrical field values in all the extension of the regions of interest, in such a way that they can be rapidly consulted, analysed and utilized in health studies or in pre-feasibility projects. Particular cases that must be considered are the electromagnetic wave propagation model and the electric characteristics of the terrain.

2 DIGITAL MAP MODELLING

To be able to model a surface and generate its Digital Terrain Model (DTM), first it is needed to obtain all the necessary data that these models will later store. This data must be easy to handle and must constitute a realistic description of the surface of interest.

The following are the selected data structure models and how they were used for the DTM implementation:

2.1 Triangular Irregular Networks

Triangular Irregular Networks (TINs) are a digital mechanism to represent surfaces given a set of measurement points. The method is based on a network of triangles that do not need to be of the same size, these triangles are joined at their vertices, called nodes, which are the points originally obtained for the definition of the terrain [4]. TINs can be generated from several kinds of sources such as data compilations taken with photogrammetric instruments, data measured on the terrain itself, or digitalized contour lines as the ones seen in Figure 1.

For the DTM implementation, digitalized contour lines were chosen as the source for the TINs because the available maps were already in this format, with the additional advantage of being also geo-referenced, an aspect that enriches the resulting DTM.

2.2 Regular Matrices

Regular matrices are based on the overlapping of a grid on the terrain and the extraction, for each cell of the grid, of the surface value on the terrain.

Given this definition, regular matrices are a raster type of structure where the cell constitutes the basic form of representation for the spatial location of the data, which is analogous to the color values of the pixels in an image [4]. Figure 2 shows an example of this
representation, where a raster has been used with a height value for each cell and a color code has been assigned to each range of height values, resulting in a smoothed colored representation of the terrain surface.

**Figure 2:** Raster structure for San Ignacio, Chaltenando, El Salvador.

### 2.3 Construction of the Digital Terrain Model

Both techniques previously described, the TINs and the regular matrices, are combined. First, given a geo-referenced map of the terrain of interest, its contour lines are generated and used to construct the TIN; then, the map is used to generate a raster structure that is overlapped on the TIN to obtain a mixed representation. This process is illustrated in Figure 3.

This mixed representation is denominated the Digital Terrain Model [5], and has the precise representation of surface irregularities of the TINs, and the easy analysis capabilities of the raster structure, both very important characteristics for the manipulation of the model in numeric and simulation processes.

**Figure 3:** DTM generation.

### 3 WAVE PROPAGATION MODELLING

A first implementation was developed to validate the construction of the DTMs using a Two Ray method [6] due to its application simplicity. Once the digital maps had been validated, they were ready to be used with a second and more robust implementation using another method able to handle more complex and realistic problems.

The Ray Tracing method [7] was chosen to develop this second implementation and its
related software tools because it supports the particularities of the problem: operation on the VHF and UHF bands of frequency, indoor and outdoor propagation, and the consideration of reflection, diffraction and refraction.

3.1 The Ray Tracing method

The Ray Tracing method is considered a high frequency method, but it does not have a lower limit in which the method is applicable, as it depends on the size of the objects in the propagation environment in comparison with the wave length of the signal; however, it is recommended to use it with frequencies larger than 100MHz [7].

The main characteristics of the method are as follows [8]:
- The method relies on the Geometric Theory of Diffraction, which states that if the frequency or the radio wave is high enough, then its propagation can be considered as it were light, orthogonal to its wave fronts.
- The method can predict the electrical field intensity at a given point by summing all the rays that arrive at said point (direct rays, reflected rays and diffracted rays).
- The method can model very precisely the propagation environment (rural or urban) by means of the simulation, discretization and processing of digital maps.
- The method can handle any radiation power and polarization as long as the radiation pattern of the antenna is modelled appropriately.

Since mobile phone frequencies are contained in the VHF and UHF bands, they are considered tropospheric propagation environments. Based on this, the Ray Tracing method can work with the following assumptions [9]:
- The propagation medium can be assumed non-magnetic. Therefore, with permeability of 1.
- The variation of the relative permittivity of the troposphere can be considered as very slow as the height increases (less than 1% for a height of 2Km), so it can be assumed constant and homogeneous in the propagation environment.
- The refraction index of the atmosphere is directly proportional to the square root of its relative permittivity, so it can also be considered constant and homogeneous.
- The UHF and VHF bands are known to be immune to the attenuation caused by rain, fog and snow, so it is acceptable to consider it a lossless propagation medium.
- Air is a substance which permittivity and permeability can be considered equal to those of the vacuum, which means, they are independent of the frequency, the electrical field magnitude and the magnetic field magnitude, so it can assumed as a linear propagation medium.

With the previous considerations, the Maxwell equations for a monochromatic wave can be expressed as:

\[
\begin{align*}
\nabla \times \mathbf{E}(r, \omega) + j \omega \mu \mathbf{H}(r, \omega) &= 0 \\
\n\nabla \times \mathbf{H}(r, \omega) - j \omega \epsilon \mathbf{E}(r, \omega) &= 0 \\
\n\nabla (\mathbf{E}(r, \omega)) &= 0 \\
\n\n\nabla (\mathbf{H}(r, \omega)) &= 0
\end{align*}
\] (1)
For a high frequency electromagnetic field, a solution in form of asymptotic series can be assumed for the equations, which then take the following form:

\[ E(r, \omega) \sim e^{-j\Psi(r)} \sum_{n=0}^{\infty} \frac{E_n}{(j\omega)^n} \]

\[ H(r, \omega) \sim e^{-j\Psi(r)} \sum_{n=0}^{\infty} \frac{H_n}{(j\omega)^n} \]

The function \( \Psi(r) \) is known as the phase function, and it depends on the form of the wave front \([10]\).

The previous expressions are known as the Lunenberg-Kline expansion for electrical and magnetic fields, commonly referred to as Ansatz. It is expected in this expansion, that the small powers are the terms giving the major contribution to the approximate solution; on the limit, when the frequency becomes infinite, the only term necessary will be the first one, when \( n = 0 \), this is known as the geometric optic field \([11]\).

For the implementation of the method, a half wave dipole antenna will be used due to its ease of application, and because its far field classifies as a geometric optic field since it can be considered as an expansion of (2) for values of \( n = 0 \). Then, the electrical field can be modelled as a series of rays which depart from the source and interact which the terrain of the map as follows \([12]\):

\[ E_{\theta} = \frac{\mu_0 \epsilon_0 e^{-j\omega} \cos\left(\frac{\pi}{2} \cos(\theta)\right)}{2\pi r} \]

The propagation medium will be represented as a tridimensional object where only the surface of the terrain will be relevant, making use of carefully constructed DTMs, as explained in section 2.

To take advantage of the triangles present in the TIN of the digital map, it is necessary and practical to use an efficient method to describe certain characteristics of its internal points, the selected method being the barycentric coordinates, as seen in Figure 4.

\[ u = \frac{d_1}{d_1 + d_2 + d_3} \]

\[ v = \frac{d_2}{d_1 + d_2 + d_3} \]

\[ w = \frac{d_3}{d_1 + d_2 + d_3} \]

**Figure 4:** Barycentric coordinates as a proportion of areas.

The Ray Tracing method identifies within the propagation medium certain zones of
interest. The Direct Visibility Zone is the first one, it is the region of the surface that has a line of sight (LOS) with the antenna. This zone is also known as the “LOS zone” or the “Illuminated zone”, Figure 5 shows an example of the application of the concept: an antenna, marked in red, is located inside the region, and the Direct Visibility Zone is marked in yellow, everything colored in green has no direct contact with the antenna or it is a region being blocked by another part of the terrain.

![Figure 5](image)

**Figure 5:** (a) Illuminated zone. (b) From another angle.

The Direct Visibility Zone is also very important region for having the potential to allow the electrical field to take high values of intensity, and because it is the zone that gives origin to the reflections [13].

The Ray Tracing method also makes use of intersection methods, which allow to establish certain visualization characteristics from a reference point known as the “observer”. Particularly, the Möller and Trumbore intersection algorithm [13] is utilized; this algorithm constructs and applies a transformation to the ray origin $O$ pointing in the direction $D$ towards the triangle defined by the vertices $p_0, p_1$ and $p_3$, as seen in Figure 6.

![Figure 6](image)

**Figure 6:** Diagram for the Möller and Trumbore intersection algorithm.

From this, the distance $t$ that exists between $O$ and the intersection point is obtained, as well as the barycentric coordinates $(u, v)$ of the intersection point. The algorithm can be summarized as the following system of equations:

$$
\begin{bmatrix}
-D_x & p_{1x} - p_{0x} & p_{2x} - p_{0x} \\
-D_y & p_{1y} - p_{0y} & p_{2y} - p_{0y} \\
-D_z & p_{1z} - p_{0z} & p_{2z} - p_{0z}
\end{bmatrix}
\begin{bmatrix}
t \\
u \\
v
\end{bmatrix}
= \begin{bmatrix}
O_x - p_{0x} \\
O_y - p_{0y} \\
O_z - p_{0z}
\end{bmatrix}
$$

(4)
If the triangle is intersected, the following inequalities must be satisfied:

\[ u \geq 0, v \geq 0, u + v \leq 1 ; \quad t \geq 0 \]  

(5)

If these conditions are not satisfied simultaneously, then it is concluded that there is no intersection within the triangle in the direction of the ray. Using this method, it can be established which elements are in LOS with the antenna analyzing the proximity of the source of the ray with the intersection point [13].

The method also needs to calculate the trajectory followed by the rays, which depends on the reflections, which in turn obey the geometry of the surface. The rays follow this trajectory until they are lost in space.

To be able to determine the trajectories it is needed to determine the direction of the first reflected ray and to verify if this reflected ray intersect a face in the mesh of the digital map. If there is an intersection, a new reflection is calculated and the previous step is repeated as many times as necessary until there is no intersection anywhere in the mesh; and finally, the set of sub-trajectories are connected to reconstruct the total path of the ray, Figure 7 a final trajectory for a ray, marked in red, after summing the contributions of individual reflections.

The starting point for the calculation of the electrical field is (3), which takes into consideration the contribution of the direct rays using a half wave dipole antenna. It is important to notice how the expression is dependent on the distance, and that this variable affects both the amplitude and the phase of the wave.

For the reflected rays, their contribution to the electrical field can be determined from the Fresnel equations, where (6) corresponds to the calculation of the complex reflection coefficient for a wave that is received with parallel polarization, and (7) for a wave that is received with orthogonal polarization.

\[
\rho_{\text{perp}} = \frac{\cos \theta - \sqrt{(e_c - j\omega)} - \sin^2 \theta}{(e_c - j\omega) \cos \theta + \sqrt{(e_c - j\omega)} - \sin^2 \theta} 
\]

(6)

\[
\rho_{\text{perp}} = \frac{\cos \theta - \sqrt{(e_c - j\omega)} - \sin^2 \theta}{\cos \theta + \sqrt{(e_c - j\omega)} - \sin^2 \theta} 
\]

(7)
Where:
- $\theta$ is the angle taken between the normal and the incident ray.
- $\varepsilon_r$ is the permittivity relative to the reflective material.
- $\sigma$ is the conductivity of the reflective material expressed in S/m.
- $f$ is the operation frequency.
- $X$ is defined as:

$$X = \frac{18 \times 10^6 \sigma}{f}$$

(8)

The polarization of the wave, which is a general case, will hardly coincide with the previous expressions; therefore, the incident electrical field will have to be decomposed in a parallel and an orthogonal component before being reflected, working separately as (9) and (10) show:

$$E_{i\|} = \left( \frac{i \times \hat{n}}{f \times \hat{n}} \right) E_i$$

(9)

$$E_{i\perp} = \overline{E_i} - E_{i\|}$$

(10)

As the last step, the reflected electrical field can be determined from (11):

$$E_r = E_{r\|} + E_{r\perp} = E_{r\|} \rho_{pep} + E_{r\perp} \rho_{orr}$$

(11)

The total field received in each triangular element of the TIN mesh in the digital map, is obtained by summing all the contributions that arrived at that particular element.

4 RESULTS

4.1 Two Ray method implementation

The Two Ray method was implemented as a problem type for the pre and post processor software GiD.

GiD manages problem types as a set of plain text files used for the configuration of a numerical model, defining variables and specific parameters to be analyzed, and defining calculations to be performed, determining with this all the possible materials and conditions that can be applied to a geometry and the format for the results that will be presented [14].

Using geo-referenced maps from a small region in the rural area of San Salvador, El Salvador, a DTM was constructed and imported to GiD to generate a triangular mesh as seen in Figure 8.
An arbitrary location for the antenna was chosen, and the problem type was executed to perform the simulation of the propagation of the electrical field, as can be seen in Figure 9. These results were compared to the measurements taken by SIGET, and it was concluded that it was a satisfactory first step of simulation, that the DTM implementation was ready to be used with a more robust implementation, and that the project could continue to its next stage.

4.2 Ray Tracing method implementation

To implement this method, a map of the natural region Cerro Verde to the west of El Salvador was employed to generate the DTMs, also using a half wave dipole antenna, with a power of 100W, an operation frequency of 800Hz and at 30m of height above the terrain.

The antenna was placed in the following coordinates: latitude of 13.81286, and longitude of -89.54453. The conductivity and permittivity relative to a moderately dry ground with a frequency of 800Hz are 2.7e-2 S/m and 15 respectively [14].

The method was implemented in Matlab using these parameters, and plugged in as a problem type in GiD to perform the simulation. Figure 10 shows the distribution of the electrical field throughout the region.
To establish a validation parameter for this implementation of the Ray Tracing method, values determined by the software ‘Radio Mobile’ were used as reference, setting as input the same conditions and values previously described. Radio Mobile is a free software application widely used in academic studies related to the calculation of electromagnetic propagation variables generated by communication systems.

Table 1 shows a sample of the data obtained with the Ray Tracing implementation, together with the data calculated by Radio Mobile, where it can be observed at first glance that very similar values for the electrical field have been calculated in areas that differ in no more than 100m.

Table 1: Sample comparison between the Ray Tracing implementation and the Radio Mobile software.

<table>
<thead>
<tr>
<th>Ray Tracing</th>
<th></th>
<th></th>
<th></th>
<th>Radio Mobile</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X</td>
<td>Y</td>
<td>dBuV/m</td>
<td>X</td>
<td>Y</td>
<td>dBuV/m</td>
</tr>
<tr>
<td>223026.15</td>
<td>1514729.49</td>
<td>77.3</td>
<td>223022.15</td>
<td>1514703.01</td>
<td>78.2</td>
<td></td>
</tr>
<tr>
<td>221949.23</td>
<td>1517570.76</td>
<td>79.9</td>
<td>222002.89</td>
<td>1517489.18</td>
<td>76.3</td>
<td></td>
</tr>
<tr>
<td>226256.91</td>
<td>1520041.42</td>
<td>87.3</td>
<td>226220.43</td>
<td>1520046.11</td>
<td>86.0</td>
<td></td>
</tr>
<tr>
<td>222427.86</td>
<td>1515058.91</td>
<td>77.5</td>
<td>222499.83</td>
<td>1515056.15</td>
<td>78.8</td>
<td></td>
</tr>
<tr>
<td>227692.80</td>
<td>1519300.22</td>
<td>85.7</td>
<td>227609.04</td>
<td>1519337.48</td>
<td>81.8</td>
<td></td>
</tr>
</tbody>
</table>

A comparison for the totality of the results can be seen in Figure 11.

Figure 8: Results comparison. Ray Tracing implementation in blue, and Radio Mobile in red.
A total of 22,531 values have been compared, obtaining an average error of 1.82dB with a standard deviation of 2.72dB. Figure 12 shows the comparison after applying an average mobile filter.

Figure 9: Filtered results comparison. Ray Tracing in blue and Radio Mobile in red.

It can be observed that there is a very good similitude between both sets of results, and it can also be evidenced how sensible the information is to the digital maps employed.

The comparison shows that within all the terrain in study there is a satisfactory matching between the values predicted by the Ray Tracing method and the values calculated by the software, which can be verified with the small average error and the small deviation value, showing that the data is not scattered and therefore more significant.

5 CONCLUSIONS

- The Ray Tracing method was able to successfully predict with satisfactory precision the values for the electrical field intensity in the direct visibility zones and in the zones affected by the field reflections, with a very acceptable average error value.
- The Ray Tracing method proved its versatility in its application to varied propagation environments and in its power to find a solution to the problem at hand.
- The Digital Terrain Models provided a very complete representation of terrain information and its characteristics, aiding with the precision of the simulation results for the electrical field values.
- The very abundant and detailed description of the propagation environment also proved to be of high relevance to improve the approximation of the results and to increase the time efficiency of the calculations.
- The implementation is ready to be improved for its use by the SIGET and other related and interested organizations.

REFERENCES

[2] SIGET. Formulación de Proyecto: Aspectos Técnicos Relativos a las Emisiones


APPLICATION OF NUMERICAL SIMULATION TO REDUCE THE FLASH OF A HOT FORGED AUTOMOTIVE PIECE

G. ABATE*, D. PEREZ, F. RIU, D. MARTINEZ KRAHMER

* Instituto Nacional de Tecnología Industrial, INTI-Mecánica
Avenida General Paz 5445, (1650) Miguelete, Buenos Aires, Argentina

Key words: Hot forging, FEM, steel, spare parts, technical assistance.

Abstract. Through of an extensive background research on the application of numerical simulation forging process, we have found that works is mainly aimed at: analyze and resolve defects [1]; improve the design of the toolkit used in forming [2,3]; predicting the microstructure of the material in the different phases of forging process [4]; microforming simulation [4]; die wear evaluation [5]; or determining lateral forces in forging dies with close warped surfaces [6].

The National Institute of Industrial Technology, through the Forging Group from INTI-Mecánica, has proposed to collaborate by introducing the use of this technology in local industries, considering the context in which they operate.

Consequently, in this work of industrial application, were analyzed using FEM software named Simufact.Forming, possible improvements about process of hot forging of a piece, that not requiring investment in machinery or in die design, and not increase production times.

To achieve this, an automotive piece was selected from various evaluated cases, basically for two reasons: because it had an excessive flash material, and the volume of production was significant for those who produced it.

Firstly, was reproduced the process as was usually the company done by simulation. Although, we worked on various improvement alternatives, especially on the size and shape of the starting material.

With a simple modification, the proposed process, produced a material savings estimated at 1.8 t / year. This modification will become an economical benefit to the company, such as for the production of several thousand additional pieces with the same quantity of material. This improvement was obtained without additional investment. Finally, the results obtained from the numerical simulation were positively confirmed by experimental tests on the forging plant floor.
1 INTRODUCTION

When is necessary to develop a piece, and is related to security, technicians and engineers aim to design and manufacture, considered the forging process as the first alternative. That is for several reasons:

1) A forged piece, has greater mechanical strength to withstand dynamic impacts or solicitations;
2) The plastic deformation process eliminates the porosity produces homogenization and a fibrous structure. Also reduces the grain size from the molten condition, increasing the toughness of the material.

What happens when you compare the benefits of a forged piece against a casting one with the same geometry? The forging process increases the mechanical strength due to the "fiberizing" of material. In other words, a forging piece compared with a casting one, designed to fulfill the same function, can be manufactured with less volume, but achieving the same strength. This is a further advantage, because material saving is significant.

The “hot forging process” is an intermediate manufacturing process, which aims to form similar geometries to the final piece, which is finished by machining, and are used in the following industries: automotive, railway, mining, energy, oil, or health (medical prostheses).

Especially in the case of hot forging, the material temperature aims to reduce the mechanical strength and increase fluidity along the forming process.

The forging process is performed at stages or steps. In the first instance the material with a very simple geometry (circular or square bars) is cutting using specific cutting weight criteria. Then, the material is heated up at the forging temperature point. The forming process is performed at one or more steps, and finally the perimeter excess material (flash) is trimmed.

In many cases, after cutting, the process begins with a pre-forming step. The purpose of this function is to distribute the material to areas where the piece will require greater volume. This step is critical in parts with large variations of section along the volume; and, if it is not used, the starting material (circular bars or square sectioned billets) is exposed at a strong deformation in one step.

Traditionally, the use of a pre-forming step, as a first step of the forming process, is up to
In order to assist the forging industry in Argentina, the Forge Group from INTI-Mechanics, who belongs to the Laboratory of Machining and Forming Processes, was established; and it began its work in 2009. In 2012 was released the first book, "Práctica Industrial de la Forja en Caliente" [8], and since then, among other activities, the Forge Group is trained constantly in the use of “Simufact.Forming”, a forming processes simulation tool [9], to publicize the software skills to the industry.

The development of a hot forged piece, involves tooling with high mechanical strength requirements and high temperature wear resistance. By the other hand, expensive machinery is required and setup times are usually extended.

This situation has encouraged the simulation tools development and application, which allows to estimate the behavior of shaping material, collaborating in the parts development, without using the expensive system of trial and error. Thus, an efficient process is achieved, in order to obtain quality parts (as tolerance, and without defects), and also a correct use of the main forging equipment, avoiding overload in the machinery and tooling.

Particularly in this activity, the main objective was to improve productivity in the use of materials, over a spare parts manufacturer. The company established the following conditions:

A. Work on parts already developed.
B. Make changes that do not require substantial investments on tooling, or on the incorporation of new machinery.
C. Keep the production rate.

2 EXPERIMENTAL PROCEDURE

2.1 Part selection analysis

The company produces over 400 different models of parts; each one of them is identified by a numeric code. In order to produce the greatest impact, the code with the highest number of units produced in 2013 was selected.

![Figure 2: Production level to the different codes](image)
From the analysis of Figure 2, we decided to analyze the forging process made of the 103 code, which has the higher production level in 2013, and through the use of simulation software, propose alternatives for improvement.

![Figure 3:Forging product code selected for analysis](image1)

### 2.2 Recognition of forging process conducted by the company (original)

It was observed that the piece was made in two forging steps (pre-final and final). However, the cavities of these dies were similar, as can be seen in Figure 4.

![Figure 4:Pre-final (left) and final upper forging die](image2)

The original process uses as starting material a bar with a circular cross section to 31.8 mm in diameter and 178 mm in length, resulting in a cutting weight of about 1100g. An image of the starting material, and the shaped piece without trimming, can be seen in Figure 5.

![Figure 5:Starting material (left) and two forged pieces with flash (right)](image3)
Analyzing the previous right image, it is evident that the flash containing the material to be subsequently discarded, it is excessive. Specifically, this scrap is 94% of the weight of the piece without trimming.

2.3 FEA Simulation of the original Process

To make changes on the forming process based on testing over the simulation software was necessary in the first instance; verify that the simulation reproduces the original process. For this, were simulated the process steps performed in the plant, and compared with the experimental results (mainly the resulting geometry).

The data required by the software for playback of the process are:

- The geometry of upper and lower dies.
- Starting material geometry, materials, and heating temperature.
- Friction coefficient between the workpiece and dies.
- Preheating of the upper and lower dies.
- Characteristics of crankshaft press (stroke, rod length, etc.).
- Sequence and duration of the several steps.

The geometry of the matrices was given to us in stl format, while the data used in the simulation are found in Table 1.

Table 1: Input data to the simulator

<table>
<thead>
<tr>
<th>Starting Material</th>
<th>Steel</th>
<th>SAE 1038</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geometry</td>
<td>Cylindrical bar</td>
<td>Φ31.8mmx178mm</td>
</tr>
<tr>
<td>Temperature</td>
<td>1200 °C</td>
<td></td>
</tr>
<tr>
<td>Friction coefficient</td>
<td>0.3 (according to reference values) [10]</td>
<td></td>
</tr>
<tr>
<td>Preheating of dies</td>
<td>180 °C</td>
<td></td>
</tr>
<tr>
<td>Press</td>
<td>Crankshaft press 1000 t</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Stroke</td>
<td>230 mm</td>
</tr>
<tr>
<td></td>
<td>Rod length</td>
<td>842 mm</td>
</tr>
<tr>
<td></td>
<td>Revolution</td>
<td>100 RPM</td>
</tr>
<tr>
<td>Steps</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

2.4 Proposals made by the Forge Group:

2.4.1 Evaluation of the need for a pre-forming step

As discussed before, when a piece to be forged has considerable variations in section along its volume, an intermediate step called pre-form is often used.

Traditionally, the use or not use of this step is a decision of the designer of the forging process, ergo, it is a practice based on previous experiences with pieces with similar
geometry.

However, investigating the state of the art on this subject, and in order to give it a theoretical foundation, we have found some publications of R. Radev [11,12,13], which proposes to establish whether or not to use a pre-form step, by using a mathematical equation that relates the geometry of the starting material, with the end piece [11].

The process applied to our generic piece in our research, the formulas and variables to consider are listed below.

Calculation for the preform estimation, applied to the study piece:

**Figure 6:** Geometric parameters required to estimate the preform over the study piece

\[
V_{ad} = V_0 - V_m \\
H_{av} = H_m + \sum \frac{A_i H_i}{A_F} \\
K_1 = \frac{V_{ad}}{V_0} \\
\phi_H = \ln \frac{H_0}{H_{av}}
\]

**Definitions:**
- \(V_0\) = Volume of the starting material
- \(V_m\) = Average volume of the piece (corresponding to \(H_m\))
- \(H_{av}\) = Average height of the part. It is a parameter which aims to connect the compressed volume, with fill volumes in the die cavity during forging.
- \(K_1\) = Dimensionless squeezed volume
- \(\phi_H\) = Vertical strain degree
According to this, if $K_1 > \varphi_h$, it will be necessary to use a pre-form, as a previous step.

### 2.4.2 1st Proposal Simulation

The first proposal is to use a starting material, also of circular section, with smaller diameter and length, to complete the piece (filling) and reduce scrap, without using a pre-form as an intermediate step.

Moreover, the idea is to perform an individual piece in a single forging step instead of two parts simultaneously in two steps, in order to keep the same production rate of the company.

It is important to mention that the proposal was to use the same die in which the company currently performs the final step (as can be seen in Figure 7), not to incur incremental costs resulting from manufacture a new tooling.

<table>
<thead>
<tr>
<th>Table 2: Input data to the simulator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Starting material</td>
</tr>
<tr>
<td>Steps</td>
</tr>
</tbody>
</table>

![Figure 7: Layout 1st starting material proposal](image)

### 2.4.3 2nd Proposal Simulation

The second proposal was to use a square sectioned material, as a starting point. This geometry allows a greater volume per unit length, compared with a circular section bar. In this newer proposal, there is not a pre-form used as an intermediate step.

As in the previous proposal, an individual piece is realized in one step forging, in order to maintain the production rate of the company in this piece.

This method is also evaluated on the die in which the company currently performs the final step (see Figure 8).

<table>
<thead>
<tr>
<th>Table 3: Input data to the simulator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Starting material</td>
</tr>
<tr>
<td>Steps</td>
</tr>
</tbody>
</table>
3 RESULTS AND DISCUSSION

3.1 Original process simulation result

The original simulation process (formed in two blows by the crankshaft press, the first one in the pre-final die and the last one in the final die) shows a good relation with the workpiece made by the company. From this situation, it is proved that the parameters used for the initial setup were appropriate, even in friction coefficient terms [10].

3.2 Estimating the need for a pre-form step

Applying the methodology of Radev to our study piece, the following conclusions can be drawn:

<table>
<thead>
<tr>
<th>Table 4: Values for estimating the need of pre-form</th>
</tr>
</thead>
<tbody>
<tr>
<td>K₁ value</td>
</tr>
<tr>
<td>0.338</td>
</tr>
</tbody>
</table>
Due to the coefficient $K_1 > \varphi_h$ it is necessary a pre-form as an intermediate step.

Nevertheless, as the company does not own the necessary equipment in order to conform a pre-form and also has not the financial possibility to purchase one. As a result, it was proposed to specifically made improvements related to the material saving, through the resizing, and the geometry of the starting material as well.

### 3.3 Results of simulation 1st Proposal

The starting material was a cylindrical bar of $\varnothing 23$ mm and 110 mm length, in accordance to the traditional calculus methodologies.

![Figure 10](image1.png) **Figure 10**: First proposal simulation (bar measures: $\varnothing 23$mm x 110 mm)

Through simulation it can be observed that the starting material it is not enough to conform the analyzed piece, because it does not fill the die, fundamentally in the “cup” sector. However, and in order to evaluate the interrelation between the simulation results and the experiments obtained in plant, we proceed to make a test in the factory. The comparison between both methods is presented in Figure 11.

![Figure 11](image2.png) **Figure 11**: Test in plant (left) & Simulation
3.4 Simulation Results 2nd Proposal

The starting material reciprocate to a square section billet of 25,4 mm side and 95 mm length, in accordance to the traditional methodologies.

The simulation results show that, the starting material would be enough to conform of the analyzed piece.

However, and in order to evaluate the interrelation between the simulation results and the experiments obtained in plant, we proceed to make a test in the factory. The comparison can be observed in Figures 13 and 14.

To sum up, the comparison between the original process and the 2nd proposal (using a square sectioned billet), the following table is shown:
Table 5: Original process characteristics vs. Proposal process

<table>
<thead>
<tr>
<th>Original process</th>
<th>2nd Proposal (square sectioned billet)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Piece formed in 2 steps (2 pieces in 2 blows).</td>
<td>One piece per blow.</td>
</tr>
<tr>
<td>Circular bar of Ø31.8 mm x 178 mm length.</td>
<td>Square sectioned billet of 25.4 mm x 95 mm</td>
</tr>
<tr>
<td>Starting material of 1100 g of cutting weight.</td>
<td>Starting material of 480 g of cutting weight.</td>
</tr>
<tr>
<td>94% of scrap.</td>
<td>66% of scrap.</td>
</tr>
</tbody>
</table>

By the proposed process, it would be achieved savings in the cutting weight of 14%, the equivalent of save around 1.8 t of steel per year, just in this analyzed piece.

By the other hand, forging a piece in a single step it is translated in a load of 550 t, which represents about a 50% of the nominal charge of the mechanical press, actually used by the company. The process will guarantee that there will not be a premature wear and/or machine braking due to an overload.

4 CONCLUSIONS

- The FEA of the hot forging process has been demonstrated a good correlation with the obtained data at the forge plant.
- In this work, the simulation of a single piece that was already developed showed that there is an improvement frame through the virtual reproduction of new processes, without taking any risk of tooling damage, or either new costs due to time of the machine dedicated of trial and error process.
- It is evident that, in order to increase the efficiency of the piece manufacturing process, it would be necessary to conform of a pre-form as a previous step to the final forging stage. This would require the company invest in an auxiliary machine, particularly, rolling forge machine
- The possibility to calculate the forging strain in a virtual way allows to take care of the involved machine, in order to avoid overloads that could lead to structural and precision failures. This kind of failures has a great cost of reparation, in addition to the lost profit due to the stopped machine.

REFERENCES


**ACKNOWLEDGEMENTS**

To Fabricio Matioli, Luis Baretta y Julián Rosso for all the support and cooperation extended during the development of this work.
Saldanha, R. R. ........................................... 1344, 1578
Salinas, L ................................................. 963
Sampaio, R ................................................. 1286, 1298, 1310
Sansinena, M .............................................. 111
Santillan Marcus, E. A .................................. 859
Santos, C. C. G ............................................. 599
Santos, J. E .................................................. 1554
Santos, M. V .................................................. 111
Santos da Silva, J. G .................................... 531
Sanziel, M. C ................................................ 859
Schaerer, C. E .............................................. 902
Scheer, S ....................................................... 996
Schiavo, L. A. C. A ....................................... 1399
Schwarz, R. A ............................................... 398
Seager, C ....................................................... 754
Senjanovic, I ................................................ 375
Shimizu, Y ...................................................... 1191
Shioya, R ....................................................... 885
Silva, B. S ...................................................... 955
Silva, D. L. C. E ........................................... 1344
Silva, J. J ......................................................... 955
Silva, V. D ....................................................... 525
Šmilauer, V .................................................... 199
Smilovich, D ............................................... 261
Soares, D ....................................................... 389
Soize, C ......................................................... 1022
Sosa, E. M ....................................................... 29
Sousa, D. F ..................................................... 281
Souza, J. R ..................................................... 1322
Spies, R. D .................................................... 1144, 1155, 1167, 1179
Staub, S ......................................................... 189
Strunin, D. V .................................................. 1103
Sumida, I. Y ................................................... 1435
Takahashi, S ................................................... 885
Takase, S ....................................................... 215
Tamagno, J. P .................................................. 1446
Tarzia, D. A ................................................... 849, 869, 876
Teixeira, J. C .................................................. 717, 1203
Teixeira, S. F .................................................. 717, 1203
Temperini, K. G ............................................. 1144
Tempone, S .................................................... 1480, 1497
Terada, K ....................................................... 215
Tinoco-Guerrero, G ....................................... 1077
Tinoco-Ruiz, J. G ......................................... 1077
Tomé, M. F .................................................... 767
Torii, A. J ..................................................... 1223
Toriu, D ........................................................ 611, 647
Tornello, M. E .............................................. 112, 122
Torres, C ....................................................... 963
Torres, I ........................................................ 705
Torres, J. A .................................................... 635
Torres, P. J .................................................... 902
Torrijos, C .................................................... 669
Toscano, B. A .............................................. 1504
Toscano, R. G ............................................... 1497
Trujillo, J ...................................................... 834
Unger, J. F .................................................... 1041
Ushijima, S ................................................... 611, 647
Valdebenito, M. A ......................................... 1257
Valencia, O ................................................... 1598
Valencia-Ramírez, A ..................................... 1077
Vaz, A. I. F .................................................... 1203
Vecchi, M. A ................................................ 1530
Velarde, P ..................................................... 919
Vicente, J ....................................................... 468
Vieira, D. A. G ............................................... 1344
Villacis, S ..................................................... 729
Vitorio Junior, P. C ........................................... 823
Vladimir, N .................................................... 375
Vrech, S. M ................................................... 225, 249
Weber, H. I ................................................... 272
Woźniak, M ................................................... 931
Wolf, W. R ................................................... 1065, 1399, 1411
Yamouni, S ................................................... 1423
Yashiro, K ..................................................... 215
Zaritzky, N ................................................... 1111
Zauli, F. P. B ............................................... 1344
This volume contains the Technical Programme of the papers accepted for presentation at the 1st Pan-American Congress on Computational Mechanics, PANACM 2015, Buenos Aires, Argentina on April 27-29, 2015.