XVIII International Conference on Water Resources
CMWR 2010
J. Carrera (Ed)
CIMNE, Barcelona 2010

A NOVEL GENETIC PROGRAMMING HEURISTIC SUPPORTED ON GENETIC ALGORITHMS AND AVERAGE MUTUAL INFORMATION AS FITNESS EVALUATION METRIC – SOME APPLICATIONS FROM HYDRAULICS –

Jaime A. Moreno[*], Eder G. Cárdenas[*] and Nelson Obregón[†]

[*] Pontificia Universidad Javeriana (PUJ)
Cra. 7 40 - 62, Bogotá, Colombia
e-mail: moreno-jaime@javeriana.edu.co, eder.cardenas@javeriana.edu.co, web page: http://www.javeriana.edu.co

[†] Instituto Geofísico Pontificia Universidad Javeriana (PUJ)
Lorenzo Uribe Building Cra. 7 42-27 floors 5 and 7, Bogotá, Colombia
e-mail: nobregon@javeriana.edu.co, web page: http://www.javeriana.edu.co

Key words: Genetic programming, average mutual information, models, objective function.

Summary. *All models aim to represent the reality of a process within the most incident variables as faithfully and easily as possible. In this order, it has developed a new genetic programming heuristic that contributes to find mathematical models that include the main variables and their relationships. This novel approach uses the average mutual information (AMI) as evaluation metric and incorporates Genetic Algorithms (GA) to find the optimal set of parameters.*
Two controlled experiments were made to assess the heuristic's behavior, they consisted in getting back to the Manning equation and Hallermeier equation from simulated data. Finally, it's shown how the algorithm achieves to recover the equations.

1 INTRODUCTION

If limitations of our senses and instrumental uncertainty of measurement, which has decreased thanks to new technologies, are ignored, the environment can be modeled and its behavior predicted to some degree of reliability. Models which could be used to this purpose can be classified into three main groups: physical models, analogue models and mathematical models, the latter being widely used due to their low cost and easy implementation in computers which let save time. The general structure of a mathematical model consists of inputs, parameters, one mathematical operator and outputs. Inputs are analogous to the processes that trigger flows of matter, energy and information in the system; parameters represent, in some models, physical characteristics of the system; and the mathematical operator is responsible for converting inputs into outputs.

One of the required steps that any genetic programming heuristic should address is the definition of one or more objective functions that are crucial in the processes of selection,

mutation and crossover. The objective functions are responsible for measuring the differences between observed values and those simulated by the model. For this reason, minimization of these functions is used in the model calibration process[2]. However, whereas in most models (based whether on neural networks, autoregressive , etc), the objective functions are used only to determine the parameters of the model, in genetic programming they also play an important role in the construction of the mathematical operator, which has no predetermined structure. This is why it is considered that the selection of the objective function or functions to be implemented should be studied for this type of tools.

In this sense, it is believed that the Average Mutual Information (AMI) of two random variables, which is a measure of the reduction of uncertainty of one of them from knowledge of the other, can highly contribute to the definition of mathematical operators and their parameterization, even though it does not have exactly the characteristics of a performance metric. This paper compares the results obtained using the AMI for seven of the commonly used performance metrics, conducting a controlled experiment that seeks to rebuild the Manning equation and Hallemeier equation.

This document is organized into six sections: the first is the introduction; the second presents a bref approach to genetic programming and genetic algorithms; the third, the seven objective functions and the AMI; in the fourth, the description of the developed genetic programming model; in the fifth, experimental design. Final remarks are presented in the sixth section.

2 THEORETICAL FRAMEWORK

2.1 An approach to genetic programming

Genetic programming is a methodology framed within the familiar artificial intelligence systems and it is based on the evolutionary theory proposed by Charles Darwin (1838). It consists in creating an initial population of equations or programs involving the possible variables that excite certain processes, which evolve from generation to generation according to principles such as reproduction, crossover and mutation[4]. Then, by introducing an objective function, the strongest individuals (equations) are selected to transfer a part of their genetic information to the next generation. It is a cyclic sequence that perfects a mathematical operator that allows modeling a process more accurately.

The product of this exercise is an operator that may describe the physical operation of the analyzed system.

Therefore, genetic programming is able to solve the inverse problem of modeling in an automated and bio-inspired manner.

2.2 Genetic algorithms

Just as genetic programming, the genetic algorithms technique is one of the five classes of systems under the name of evolutionary algorithms. Therefore, they maintain many

similarities in their functioning. Genetic algorithms have their beginnings in the seventies as a method of optimization. They start out from a set of possible solutions to a specific problem, which are codified in character strings called chromosomes which are processed again and again through features that mimic natural selection, reproduction and mutation to find the best solution for the problem[3].

The most significant difference between these two tools is that while genetic programming is responsible for finding the structure of the mathematical model, genetic algorithms are responsible for finding the parameters of the structure of a model previously established.

3 OBJECTIVE FUNCTIONS AND AVERAGE MUTUAL INFORMATION

3.1 Objective Functions

In **Table 1** the metrics used as objective functions are shown[1].

NAME	DESCRIPTION	EQUATION
Mean Absolute Error	This metric records in real units the level of overall agreement between the observed and modelled datasets.	$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left\lvert Q_i - \widehat{Q}_i\right\rvert$ (1)
Root Mean Squared Error	This metric records in real units the level of overall agreement between the observed and modelled datasets	$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(Q_i-\widehat{Q}_i)^2}{n}}$ (2)
Fourth Root Mean Quadrupled Error	This metric records in real units the level of overall agreement between the observed and modelled datasets.	$\mathrm{R4MS4E} = \sqrt[4]{\frac{\sum_{i=1}^{n}(Q_i-\widehat{Q}_i)^4}{n}}$ (3)
Relative Absolute Error	This metric comprises the total absolute error made relative to what the total absolute error would have been if forecast had simply been the mean of the observed values.	$\mathrm{RAE} = \frac{\sum_{i=1}^{n}\lvert Q_i-\widehat{Q}_i\rvert}{\sum_{i=1}^{n}\lvert Q_i-\overline{Q_i}\rvert}$ (4)
Mean Absolute Relative Error	This metric comprises the mean of the absolute error made relative to the observed record. It has also been termed "relative mean error".	$\mathrm{MARE} = \frac{1}{n}\sum_{i=1}^{n}\frac{\lvert Q_i-\widehat{Q}_i\rvert}{Q_i}$ (5)
Median Absolute Percentage Error	This metric comprises the median of the absolute error made relative to the observed record.	$\mathrm{MdAPE} = \mathrm{Median}\left(\left\lvert\frac{Q_i-\widehat{Q}_i}{Q_i}\right\rvert\right) * 100$ (6)
Mean Squared Relative Error	This metric comprises the mean of the squared relative error in which relative error is error made relative to the observed record.	$\mathrm{MSRE} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Q_i-\widehat{Q}_i}{Q_i}\right)^2$ (7)

Table 1: Evaluation metrics

It has been reported that this objective functions are suitable to deals with the fact that some parameters might be uncorrelated in describing the behavior of the system, which is proven in controlled experiments.

3.2 Average Mutual Information (AMI)

The Average Mutual Information between two random variables X_1 y X_2 is an uncertainty reduction measure of X_1 based on knowledge of X_2. It may be described as a measure of the amount of information that a random variable contains of another random variable[5]. This is defined by the following equation:

$$I(X_1, X_2) = \sum_{x_1 \in X_1} \sum_{x_2 \in X_2} p(x_1, x_2) \log \frac{p(x_1, x_2)}{p(x_1)p(x_2)} \quad (8)$$

Where:
$p(x_1, x_2)$ = Joint probability of variables X_1 and X_2
$p(x_1)$ = Marginal probability of variable X_1
$p(x_2)$ = Marginal probability of variable X_2

One of the advantages of the AMI to be used as a metric is that it does not only identify linear correlations between two variables (**Figure 1**). For this reason, the developed tool uses AMI, normalized from 0 to 1, as a performance criterion to select the structures that best fit the data.

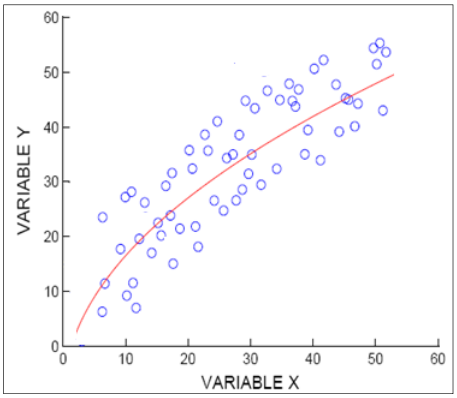

Figure 1: Nonlinear correlation

4 MODEL APPROACH

The heuristic presented, which was implemented using a Matlab ® code, is presented in the **Figure 2** and consists basically of the following steps:

Initially data is read and then organized into a table, so that the first column corresponds to the output variable and the other columns correspond to the variables that could be part of the resulting equation. Subsequently, using a set of operators, which in this case correspond to addition, subtraction, multiplication, division, exponential, natural logarithm and power function, the new variables are obtained, which are the result of applying the above functions to the input variables. Given the functions that were incorporated into the program, in some cases these require two arguments and in some other, only one. Finally, the resulting series are organized in three-dimensional arrays.

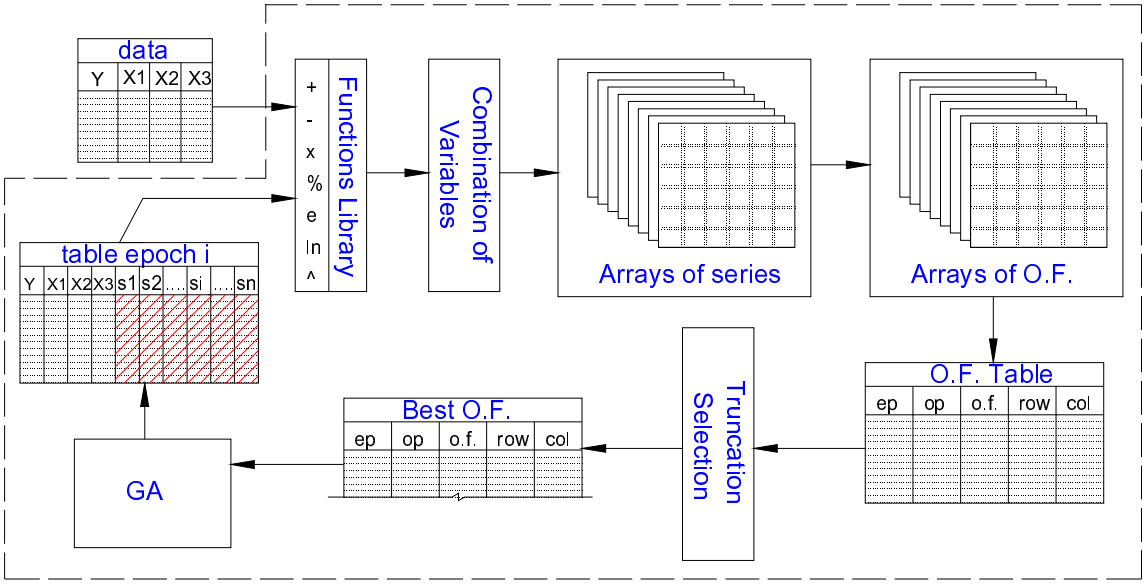

Figure 2: Model Structure

To evaluate the performance of the new individuals, each of them is compared against the output variable through an evaluation metric, which can be selected within a set of eight options, which correspond to those in **Table 1** plus the AMI. The calculated values of said metric are incorporated into a table, in which are also stored the indices of the functions that produce values. Likewise, the index of the time at which that individual was conceived is kept. Subsequently, individuals are selected based on the values of the performance metric using the mechanism known as "Truncation Selection"[6] which retains the best proportion of individuals and discards the rest.

To get maximum performance from each of the individuals selected, they undergo a process through which the best exponents of each of the composing variables are identified. To accomplish this, the genetic algorithms technique is used. However, since the function to be optimized is defined by the user, the option to choose from one of the following eight metrics: MAE, RMSE, R4MS4E, RAE, MARE, MdAPE, MSRE or the squared correlation coefficient R^2, was generated. As can be seen, the first seven correspond to the same ones used to make the selection of individuals, while the last, which used to correspond to the AMI, was replaced by the squared correlation coefficient R^2, because although the Average Mutual Information, as mentioned above, serves to identify non-linear correlations, in this case would not identify the appropriate exponents for each of the variables.

Finally, selected and refined individuals through this method are incorporated as new variables for the next cycle. The process ends when the critical value of the objective function has been got or when it meets a certain number of times has been reached.

5 EXPERIMENTAL DESIGN

In order to test the algorithm and code capacity, and to determine the best objective function, it was attempted to recover the Manning equation, which is an empirical equation, resulting from the work of several researchers[8]. It expresses the velocity versus hydraulic radius (R), the slope of the energy line (s) and a coefficient (n) representing the flow resistance when passing over certain area.

$$v = \frac{1}{n} * R^{2/3} * s^{1/2} \tag{9}$$

This is a nonlinear equation with inconsistent dimension, however, is the most widely used equation in the calculation of uniform flow in channels, due to its simplicity.

The procedure utilized to recover the equation involved the generation of random data sets of 1000 patterns, uniformly distributed in each of the independent variables, keeping the orders of magnitude characteristic of each of these (**Table 2**). Later, from the equation uniform flow velocity was calculated for each of the records, thus obtaining series of equal length.

Variable / Statistic	v (m/s)	n (-)	R (m)	s (m/m)
Mean	4.9817	0.0297	1.4886	0.0106
Standard Deviation	3.8907	0.0115	0.8581	0.0055
Minimum	0.0482	0.0100	0.0023	0.0010
Maximum	25.4029	0.0500	2.9981	0.0200
Count	1000	1000	1000	1000

Table 2 : Manning equation variables statistics

However, since the success of heuristics is subject largely to the randomness provided by genetic algorithms, 640 repetitions were performed, 80 for each of the eight objective functions, in order to assess the stability of the solutions depending on the metric used.

Because of the necessity for an indicator of the stability of the solutions obtained from the different performance metrics, it was calculated the coefficient of variation of the 80 values for the eight cases analyzed. Metrics with less variation in results were AMI and R4MS4E, the one which showed higher variability was MARE (**Table 3**). Additionally, regarding the structure of the obtained mathematical operators, the best results were obtained with AMI, because in these cases the structure of the found mathematical operator agreed with the Manning equation in 86% of the cases.

ObjFun	Mean	Best Result	Worst Result	C_V
MAE	1.63	0.39	1.74	0.13
RMSE	1.65	0.54	2.32	0.16
R4MS4E	3.34	1.58	3.60	0.08
RAE	0.56	0.12	0.60	0.17

ObjFun	Mean	Best Result	Worst Result	C_V
MARE	0.25	0.06	0.37	0.28
MdAPE	25.49	15.60	28.76	0.09
MSRE	0.21	0.06	0.23	0.11
AMI-R^2	0.84	0.95	0.72	0.08

Table 3 : Coefficients of variation of the results according to each objective function

To compare the best results obtained by different metrics, from each of the eight groups was selected the one that submitted the best value in terms of the respective performance metric. Subsequently, they were compared with each other using the seven other metrics (**Table 4**). Likewise, scatter diagrams were plotted (**Figure 3**).

Metric / ObjFun	MAE	RMSE	R4MS4E	RAE	MARE	MdAPE	MSRE	AMI
MAE	0.3372	0.5589	1.1115	0.1176	0.0946	5.0543	0.0275	0.6885
RMSE	0.6326	0.7725	1.0965	0.2206	0.1846	17.5168	0.0521	0.6680
R4MS4E	1.0231	1.2095	1.4520	0.3567	0.5570	20.4468	3.1418	0.4848
RAE	0.4043	0.5886	1.0120	0.1410	0.1158	6.9697	0.0356	0.7447
MARE	0.1828	0.3036	0.6013	0.0637	0.0500	3.5051	0.0051	0.7539
MdAPE	0.8465	1.2487	2.1110	0.2951	0.2788	13.3561	0.1898	0.6182
MSRE	0.6680	0.9240	1.4930	0.2329	0.1603	14.6388	0.0392	0.5825
AMI	0.2492	0.2860	0.3479	0.0869	0.0586	5.4920	0.0038	0.8708

Table 4 : Comparison of objective functions

The above analysis allowed determining that the best results in terms of model performance and stability of the generated solutions are presented when using the AMI as objective function for the selection of individuals and the squared correlation coefficient R^2 for the determination of the exponents.

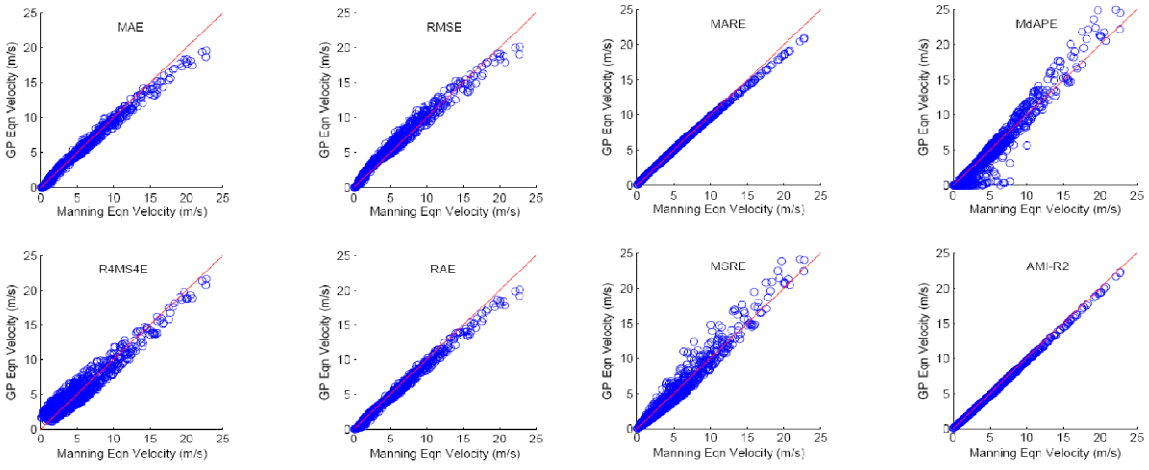

Figure 3: Comparison of evaluation metrics as objective functions

Finally, it was sought to reconstruct the Hallermeier equation[7], using the AMI and the squared coefficient of correlation as objective functions for the selection of individuals and the refinement of the exponents of the equation respectively. This equation allows calculating the settling velocity of a particle (Vs) in transitional flow, from the acceleration of gravity (g), specific weight (γ), particle diameter (d) and the kinematic viscosity of the fluid (ν).

$$v_s = (g^{0.7} * \gamma^{0.7} * d^{1.1})/\nu^{0.4} \quad (10)$$

As Manning equation experiment, random data sets of 1000 patterns were generated for the independent variables, the settling velocities values were obtained from Equation (10). **Table 5** shows Hallermeier equation variables statistics. The obtained equation (Equation 11), presents a mathematical operator of the same characteristics as the original equation and the exponents to which each of the variables are similar.

Variable Statistic	v_s (cm/s)	γ (Ton/m^3)	d (mm)	ν (cm^2/s)
Mean	34.3871	1.7053	0.0444	0.0126
Standard Deviation	18.1050	0.0198	0.0216	0.0015
Minimum	4.7136	1.6700	0.0080	0.0101
Maximum	73.0507	1.7399	0.0829	0.0152
Count	1000	1000	1000	1000

Table 5 : Hallermeier equation variables statistics

$$v_s = (g^{0.73} * \gamma^{0.66} * d^{1.09})/\nu^{0.38} \quad (11)$$

6 FINAL REMARKS

- The best results were obtained using the AMI as objective function for the selection of individuals and the squared correlation coefficient R^2 for the determination of the exponents of the variables.
- Although AMI showed a good performance criterion in the two controlled experiments, it should be tested on other PG heuristics and equations of different structures.
- The reason for the AMI's good performance criterion derives from its own definition: when it is assessing whether the observed and simulated data are similar, it does not only deals with the errors between the series, like most of the metrics, it also deals with the information gain of a variable with respect to another.
- Besides the experiments in this study, tests could be made with data measured in the laboratory or with synthetic data with incorporated noise to represent the instrumental uncertainty and the uncertainty of unmeasured variables that could contribute to the well functioning of the model and further assess the benefits of this new heuristic.
- Further work about alternative formulations of a system model, when these are expressed in terms of partial differential equations, is part of a research. It is expected to be published elsewhere.

REFERENCES

[1] C. Dawson *et al*, "HydroTest: Aweb-based toolbox of evaluation metrics for the standardaised assessment of hydrological forecasts", *Environmental Modelling & Software*, **22**, 1034-1052 (2007).

[2] P. Ogou *et al*, "Multi-objective global optimization for hydrologic models", *Journal of Hydrology*, **204**, 83-97 (1998).

[3] R. Olarte, "Herramientas para la implementación de algoritmos genéticos en ingeniería civil con énfasis en hidroinformática", Civil Engineer Thesis, *Pontificia Universidad Javeriana* (2003).

[4] S. Liong *et al*, "Genetic programming: A New paradigm in rainfall runoff modeling", *Journal of the American Water Resources Association. Engng*, **38**, 3, 705-718 (2002).

[5] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc., (2006).

[6] V. Babovic and M. Keijzer, "Genetic programming as a model induction engine", *Journal of Hydroinformatics*, **2.1**, 35-59 (2000).

[7] V. Babovic *et al*, "Generation of settling velocity equations for sand grains using genetic programming", *Hidroinformatics: proceedings of the 6th international conference*, 1631-1638 (2004).

[8] V.T. Chow, *Hidráulica de canales abiertos*, McGraw Hill, 1994.