VOL. 11, NO. 3, FEBRUARY 2016 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

INVESTIGATION INTO THE ROBUSTNESS OF EVOLUTIONARY PROGRAMMING REGRESSION FOR SEDIMENTATION STUDY

Nadiatul Adilah Ahmad Abdul Ghani
Faculty of Civil Engineering and Earth Resources, University Malaysia Pahang, Malaysia
E-Mail: nadiatul@ump.edu.my

ABSTRACT

Evolutionary Polynomial Regression (EPR) has been used to determine the total sediment load in selected rivers in Malaysia. In order to test the robustness and generalization ability of EPR modelling, the approach that is generally adopted is to test the performance of trained EPR models on an independent validation set. If such performance is adequate, the model is deemed to be robust and able to generalize. When evaluating EPR models, consideration must be given not only to their predictive accuracy but also to the interpretive ability of the models. This can be done by carrying out a sensitivity analysis that quantifies the relative importance of model inputs to the corresponding outputs. In this paper, the robustness of EPR models is investigated in a case study of predicting the total sediment load at Malaysian rivers. A procedure that tests the robustness of the predictive ability of EPR models is introduced. The results indicate that the good performance of EPR models in the data used for model calibration and validation also perform in a robust fashion over a range of data used in the model calibration phase. The results also indicate that validating EPR models using the procedure applied in this study are essential in order to investigate their robustness.

Keywords: evolutionary polynomial regression, total sediment load, robustness, prediction.

INTRODUCTION

Predicting total sediment load in rivers normally used to prevent flooding especially during heavy rains. A sedimentation process in rivers changes the shapes and pattern of riverbank. Researchers had developed a model to identify the sedimentation process for estimation of the total sediment load. Some of these models include Engelund and Hansen (1967), Graf (1971), Ackers and White (1973), Yang and Molinas (1982), Van Rijn (1986), Karim (1998) and Nagy *et al.* (2002). These models were developed based on flume data from western countries, including America and Western Europe, and have not been widely used or evaluated in other parts of the world (Sinnakaudan *et al*., 2006). Since the 1990's, some Malaysian researchers have developed models based on Malaysian conditions (e.g. (Ariffin J, 2004; Chan *et al*., 2005; Sinnakaudan *et al.*, 2006). In this paper, Evolutionary Polynomial Regression (EPR) which is a data-driven hybrid regression technique was used to develop a new model for total sediment load.

EVOLUTIONARY POLYNOMIAL REGRESSION (EPR)

EPR is developed by Giustolisi and Savic (2006). It can be defined as a non- linear global stepwise regression, providing a symbolic formula of models. . The EPR technique has been used successfully in solving several problems in civil engineering (e.g. (Savic *et al*., 2006); (Berardi *et al*., 2008); (Giustolisi *et al.*, 2008)). It constructs symbolic models by integrating the soundest features of numerical regression (Draper and Smith, 1998) with genetic programming and symbolic regression (Koza, 1992). This strategy provides the information in symbolic form expressions, as usually defined and referred to in mathematical literature (Watson, 1996). The general form of expression in EPR can be presented as follows (Giustolisi and Savic, 2006):

$$y = \sum_{j=1}^{m} F(X, f(X), a_j) + a_o \quad (1)$$

where: y is the estimated vector of output of the process; m is the number of terms of the target expression; F is a function constructed by the process; X is the matrix of input variables; f is a function defined by the user; and aj is a constant. A typical example of EPR pseudo-polynomial expression that belongs to the class of equation (1) is as follows (Giustolisi and Savic, 2006):

$$\hat{Y} = a_o + \sum_{j=i}^{m} a_j \cdot (X_1)^{ES(j,1)} \ldots\ldots (X_k)^{ES(j,k)} \cdot f\left[(X_1)^{ES(j,k+1)} \ldots\ldots (X_k)^{ES(j,2k)}\right] \quad (2)$$

where: is the vector of target values; m is the length of the expression; aj is the value of the constants; Xi is the vector(s) of the k candidate inputs; ES is the matrix of exponents; and f is a function selected by the user.

Referring from D. Laucelli *et al* (2011), EPR is a hybrid data-mining modelling technique whose main features are explicitly stated in its name. It's called Evolutionary because it employs a population based strategy for searching optimal models by mimicking the evolution of the fittest individual in nature. In particular it

employs a Genetic Algorithm (GAs) (Goldberg, 1989) to find the optimal sets of exponents in equation (2) within the combinatorial search space, as defined by the user defined set of exponents. It is Polynomial because EPR mathematical structures, e.g. equation (2) are linear with respect to their parameters although not necessarily linear in their attributes (due to both exponents different from 1 and possible selection of function f). EPR is actually a Regression technique since model parameters of any 'pyseudo-polynomial expression' are computed from data.

EPR is suitable for modelling physical phenomena, based on two features (Savic *et al*., 2006): (i) the introduction of prior knowledge about the physical system/process - to be modelled at three different times, namely before, during and after EPR modelling calibration; and (ii) the production of symbolic formulas, enabling data mining to discover patterns which describe the desired parameters. In the first EPR feature (i) above, before the construction of the EPR model, the modeller selects the relevant inputs and arranges them in a suitable format according to their physical meaning. During the EPR model construction, model structures are determined by following user-defined settings such as general polynomial structure, user-defined function types (e.g. natural logarithms, exponentials, tangential hyperbolic) and searching strategy parameters. The EPR starts from true polynomials and also allows for the development of non-polynomial expressions containing user-defined functions (e.g. natural logarithms). After EPR model calibration, an optimum model can be selected from among the series of models returned. The optimum model is selected based on the modeller's judgement, in addition to statistical performance indicators such as the coefficient of determination (CoD). A typical flow diagram of the EPR procedure is shown in Figure 2, and detailed description of the technique can be found in (Giustolisi and Savic, 2006).

CASE STUDY

338 data from the year 1999 till 2007 at 10 selected rivers in Malaysia were used to develop the EPR model. The data used for model calibration and validation were collected from the Department of Irrigation and Drainage (DID), Ministry of Natural Resources and Environment, Malaysia (hereinafter referred to as the DID). The first set of data was collected from the Pari River in Taman Merdeka and Kerayong River in Kuala Lumpur from 1998 to 1999. The second set of data was undertaken at the Kinta River catchment, which consists of four rivers including Kinta River, Raia River, Pari River and Kampar River. The third set of data took place over the period 2000 to 2002, at the Langat River catchment area, comprising Langat River, Lui River and Semenyih River. The fourth and final set of data was completed at the Kulim River in 2007.

MODEL DEVELOPMENT USING EPR

The EPR model was developed using the available software package, EPR Toolbox Version 2 (Laucelli *et al*., 2009). A set of 338 data represents the sediment transport features of ten different rivers across Malaysia were used in this study.

The first important step in the development of the EPR model was to identify the potential model inputs and outputs. Based on previous studies carried out by many researchers (e.g. Sinnakaudan, 2008), for the purpose of this study, eight inputs were utilized, having deemed them to be the most significant factors affecting the sediment transport. These inputs include the hydraulic radius (R), flow depth (Yo), flow velocity (V), median diameter of sediment load (d_{50}), stream width (B), water surface slope (So), fall velocity (ωs) and flow discharge (Q). The total sediment load (Tj) was taken as the output model.

The data division is taken as a next step in the development of the EPR model. The data were randomly divided into two sets: a training set for model calibration and an independent validation set for model verification. In dividing the data into their sets, the training and testing sets were selected to be statistically consistent, thus, represent the same statistical population, as recommended by Shahin *et al*. (2004). In total, 271 data cases (80%) of the available 338 data cases were used for training, and 67 data cases (20%) were used for validation

The following step in the development of the EPR model was selecting the related internal parameters for evolving the model. This was carried out by a trial-and-error approach in which a number of EPR models were trained, using the parameters given in Table-1, until the optimum model was obtained. A more detailed description of the modelling parameters used in Table-2 can be found in the EPR Toolbox manual (Laucelli *et al*., 2011).

Table-1. Internal parameters used in the EPR modeling.

Parameter	EPR setting
Regression type	Statistical
Polynomial structure	$Y = sum(a_i \times X_1 \times X_2 \times f(X_1) \times f(X_2)) + a_o$
Function type	Exponent
Term	[1:5]
Range of exponents	[0, 0.5, 1, 2]
Generation	10
Offset (a_o)	Yes
Constant estimation method	Least Square

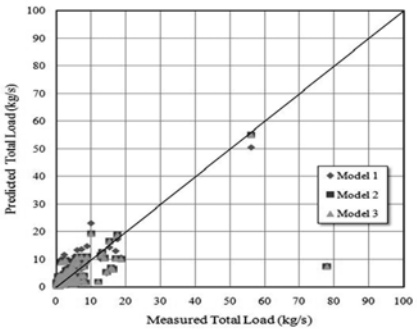
Performance indicators

The trial-and-error approach in which a number of EPR models were trained with different internal modelling parameters, gave three models with the best results, as shown in Table-2 and graphically in Figure-1(a) and 1(b). It can be seen from Figure-1(a) and Figure-1(b) that there is not a great deal of scatter around the line of equality between the measured loads and the validation set, the performance of the total load model for the three models looks similar.

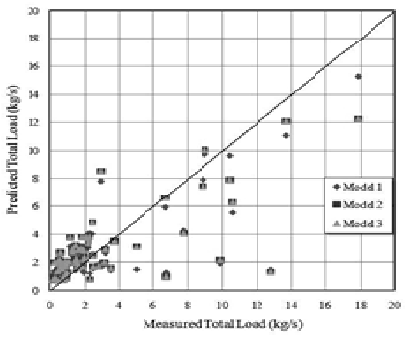
Five performance measures namely: the coefficient of correlation, *r*, coefficient of efficiency, *E*, root mean squared error, *RMSE*, discrepancy ratio, *DR*, and Akaike information criterion, *AIC* was used to evaluate the relationship between the measured and predicted total loads. The coefficient of correlation, *r*, is the performance measure that is widely used in civil engineering but sometimes can be biased in reflecting higher or lower values, leading to misleading model performance. The coefficient of efficiency, *E*, is an unbiased performance estimate and provides an assessment of the overall model performance, which can range from minus infinity to 1.0, with higher values indicating better agreement (Legates and Mc Cabe, 1999). The *RMSE* has the advantage in that large errors receive much greater attention than small errors, as indicated by Shahin *et al*. (2004). The discrepancy ratio, *DR*, as indicated by Sinnakaudan et al. (2006) is the ratio between the predicted and measured total sediment loads, and a model is considered to be suitable if its discrepancy ratio falls within the range of 0.5–2.0. The *AIC* gives an estimate of the expected relative distance between the fitted model and the unknown true model. The smallest value of *AIC* is considered to be the most favourable amongst the set of candidate models (Shaqlaih *et al*., 2011).

Table-2. Performance results of the EPR models in the training and testing sets.

Performance measurement	Model–1	Model–2	Model–3
Correlation coefficient, *r*			
Training	0.72	0.72	0.73
Validation	0.74	0.74	0.74
Coefficient of efficiency, *E*			
Training	0.52	0.52	0.52
Validation	0.55	0.55	0.55
RMSE			
Training	2.46	2.46	2.46
Validation	2.41	2.41	2.41
Discrepancy ratio, *DR*			
Training	0.68	0.69	0.69
Validation	0.64	0.66	0.66
AIC			
Training	0.00	4.10	4.00
Validation	0.00	5.20	5.20



(a)



(b)

Figure-1. Performance of the EPR model: (a) Training set; (b) Validation set.

Three best EPR models in Table-2 shows that *r*, *E*, *RMSE* and *DR* close to each other and all three models have consistent performance in both the training and testing sets. However, based on the *AIC* results, it shows that Model–1 is superior to the other models and can be considered to be optimal. As can be seen in the following equations (i.e. equation. 3–5, Model–1 has only 6 input variables, equation (3), whereas both Model-2 equation (4) and Model–3 (equation 5) have 8 input variables each. It should be noted that the performance results of these models are considered to be acceptable in representing the sediment transport problem compared to those of the most available methods, as will be seen in the next section. The symbolic formulae obtained from the EPR Models are as follows:

$$T_j = 226356.81\, V\, d_{50}^{2} + 18.37\, Q^{0.5}\, Y_o\, S_o^{0.5}\, e^{0.5V} + 0.000012\, Q\, d_{50}^{0.5}\, e^{0.5B} \quad (3)$$

$$T_j = 222250.88\, V\, d_{50}^{2} + 18.17\, Q^{0.5}\, Y_o\, S_o^{0.5}\, e^{0.5V} + 0.000012\, Q\, d_{50}^{0.5}\, e^{0.5B} + 1.23\, Q\, Y_o\, \omega_S^{2}\, R^2\, S_O\, e^{2\omega s+2R} \quad (4)$$

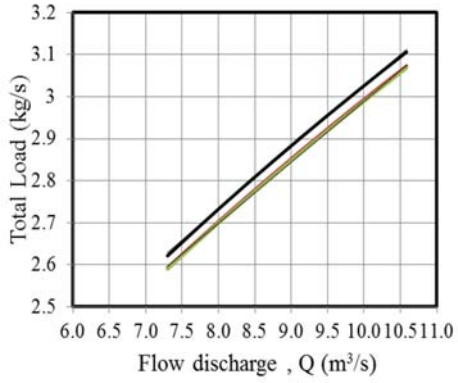
$$T_j = 162.24\, B^2\, Y_o\, \omega_s^{2}\, R^2\, S_o^{2} + 222624.92\, V\, d_{50}^{2} + 18.15\, Q^{0.5}\, Y_o\, S_o^{0.5}\, e^{0.5V} + 0.000012 Q d_{50}^{0.5} e^{0.5B} + 0.000023\, Q^2\, \omega_s\, R^2\, e^{2R} \quad (5)$$

where: T_j is the total sediment load, V is the flow velocity, d_{50} *is* the median diameter of sediment load, Q is the flow discharge, Y_o is the flow depth, S_o is the water surface slope, B is the stream width, R is the hydraulic radius and ω_s is the fall velocity.

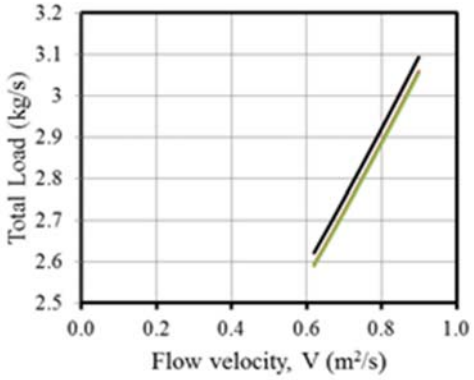
Robustness study

In order to confirm the robustness of the best EPR model (Model 1), an additional validation approach was utilized, as proposed by Shahin *et al*. (2004). The approach consists of carrying out a parametric study, part of which includes investigating the response of the EPR model output to changes in its inputs. All input variables, except one, were fixed to the mean values used for training, and a set of synthetic data (between the minimum and maximum values used for model training), was generated for the input that was not set to a fixed value. The synthetic data set was generated by increasing its values in increments equal to 5% of the total range between the minimum and maximum values, and the model response was then examined. This process was repeated using another input variable until the model response has been tested for all input variables.

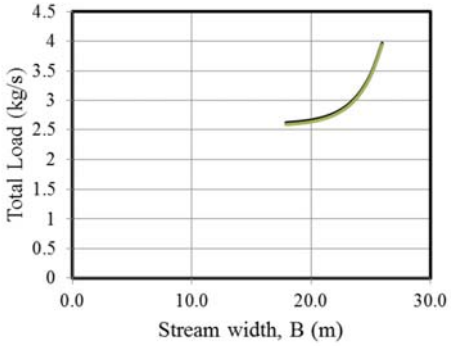
The robustness of the model were tested by examining how well the trends of the total sediment loads, over the range of the inputs examined, are in agreement with the underlying physical meaning of sediment problem. The results of the robustness study are shown in Figure-2, which agree with hypothetical expectations based on the known physical behaviour of the total sediment load. Figures-2 (a–h) shows that the predicted total sediment load increases in a relatively consistent and smooth fashion, as the discharge, velocity, width, river depth, median diameter, slope, hydraulic radius and fall velocity increase. Input parameter for Model 1 (green line) stated in Figure-2 (a-f), while input parameter for Model 2 (black line) and Model 3 (red line) stated in Figure-2(a-h).



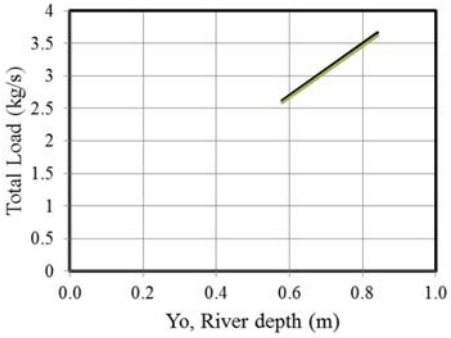
(a)



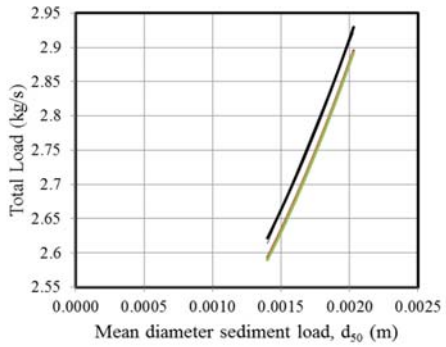
(b)



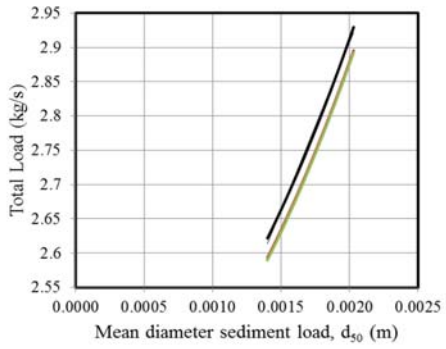
(c)



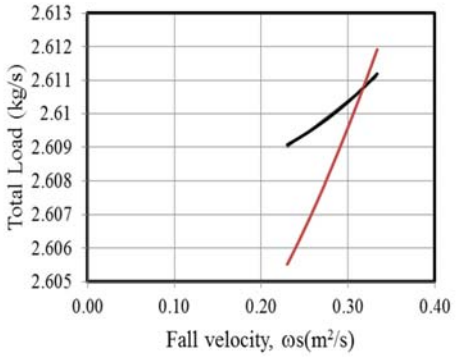
(d)



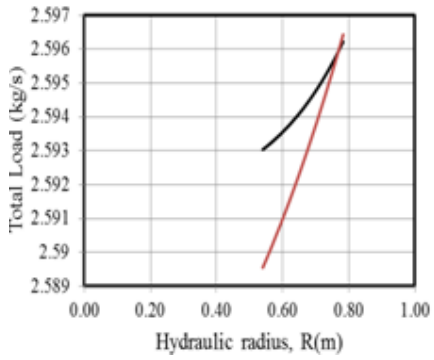
(e)



(f)



(g)



(h)

Figure-2. Robustness study showing the EPR model ability to generalise.

Sensitivity analysis

The interpretive ability of the model also been considered when evaluating the best EPR model. This can be done by carrying out a sensitivity analysis that quantifies the relative importance of model inputs to the corresponding outputs. The relative importance was determined using three different sensitivity measures, namely the range (r_a), gradient (g_a) and variance (v_a), in this study (Cortez *et al*., 2009):

$$r_a = \max(y_a) - \min(y_a) \tag{6}$$

$$v_a = \sum_{j=2}^{L} (y_{a,j} - \bar{y}_a)^2 / (L-1) \tag{7}$$

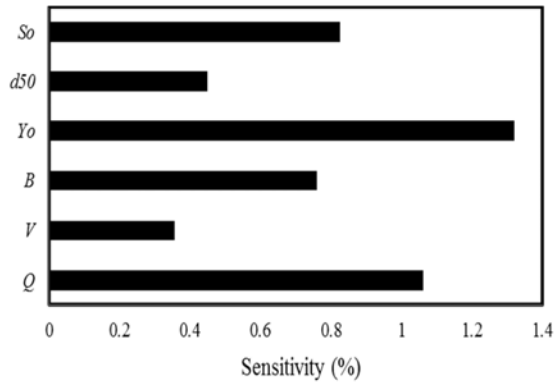
$$g_a = \sum_{j=2}^{L} \left| y_{a,j} - y_{a,j-1} \right| / (L-1) \tag{8}$$

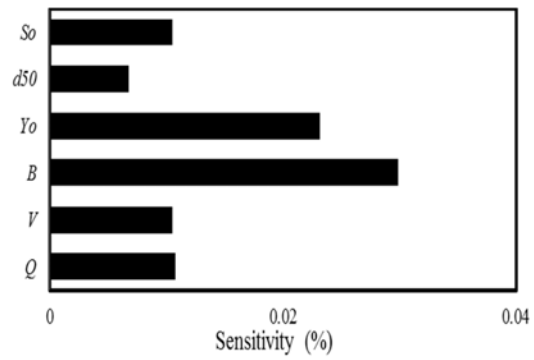
For all of the above metrics, the higher the value the more relevant is the input. Thus, the relative

importance (R_a) can be given as follows (Cortez *et al*., 2009):

$$R_a = s_a / \sum_{i=1}^{I} s_i \times 100(\%) \quad (9)$$

where: $y_{a,j}$ is the sensitivity response for $x_{a,j}$ and s is the sensitivity measure (i.e. r, g or v). Figure 3 shows the graphical representation of the relative importance measures for Model 1 in the form of bar charts. The first is simply the sensitivity according to range. Consider the sensitivity results for the model inputs in Figure-3. The relative importance model inputs for total load according to range are determined by using equation 6. A second possible measure is the variance produced in the output when the input is moved through its entire range. Equation 7 were use to determine the sensitivity according to variance. A third possible measure for sensitivity is the average gradient over all the intervals. Equation 8 used to determine the sensitivity according to the gradient measure. The results in Figure-3 show that all three measures capture the higher sensitivity of river depth, *Yo* as input variable compared to others variables. It can be seen that the river depth, Y_o, seems to provide greater importance than the other input variables for almost all sensitivity measures used, while the flow velocity, *V*, and median diameter of sediment load, d_{50}, hold less importance than the other input variables.





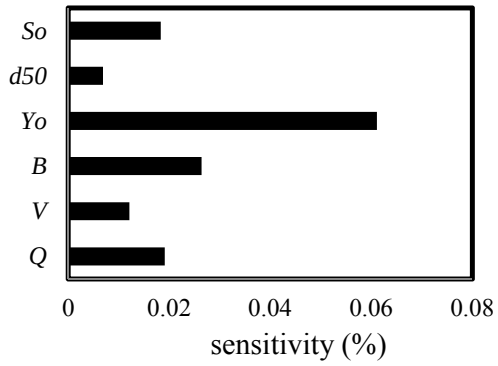


Figure-3. Sensitivity analysis showing the relative importance of the EPR model inputs.

CONCLUSIONS

Using data provided by Department of Irrigation and Drainage (DID), Ministry of Natural Resources and Environment Malaysia, new sediment transport model was develop using Evolutionary Polynomial Regression technique. From that, three EPR model were selected and had been analyses to get the best model. The performance of the three EPR model in relation to the validation set showed less scattering around the line of equality between the measured and predicted total sediment loads. The statistical analyses used for comparison included the coefficient of correlation, *r*, root mean squared error, *RMSE*, coefficient of efficiency, *E*, discrepancy ratio, DR, and Akaike information criterion, *AIC*. The results shows that EPR Model 1 is the best model with *r*, *RMSE*, *E*, *DR* and *AIC* were found to be equal to 0.74, 2.41, 0.55, 0.64 and 0.00, respectively.

The EPR Model 1 was also found to be robust in terms of its generalisation ability as its behaviour was found to be in agreement with the underlying physical meaning of sediment transport. The sensitivity analysis was also carried out to check the relative importance of model inputs to the corresponding output. The sensitivity analysis indicated that the river depth, Y_o, provided greater importance than the other input variables, while the flow velocity, *V*, and median diameter of sediment load, d_{50}, and hold less importance than the other input variables.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the Department of Irrigation and Drainage (DID), Ministry of Natural Resources and Environment, Malaysia for providing the data used in this study.

REFERENCES

F. Engelund and Hansen. 1967. A monograph on sediment transport in alluvial streams. Denmark: Copenhagen. Teknisk Forlag.

W.H Graf. 1971. Hydraulics of sediment transport. New York: McGraw Hill.

P. Ackers and W. R. White. 1973. Sediment transport: new approach and analysis. Journal of the Hydraulics Division. ASCE, vol. 99, pp. 2041-2060.

C. T. Yang and A. Molinas. 1982. Sediment transport and unit stream power function. Journal of Hydraulic Engineering, ASCE. vol. 108(6), pp. 774-793.

L.C. Van Rijn. 1986. Mathematical modelling of suspended sediment in non-uniform flows. Journal of Hydraulic Engineering, ASCE. vol. 112(6), pp. 433-455.

F. Karim. 1998. Bed material discharge prediction for nonuniform bed sediments. Journal of Hydraulic Engineering, vol. 124(6), pp. 597-604.

H.M. Nagy , K. Watanabe and M. Hirano. 2002. Prediction of sediment load concentration in rivers using artificial neural network model. Journal of Hydraulic Engineering, ASCE, vol. 128(6), pp. 558-595.

S.K. Sinnakaudan, A. Ab.Ghani, M.S. Ahmad and N.A. Zakaria. 2006. Multiple linear regression model for total bed material load prediction. Journal of Hydraulic Engineering, ASCE, vol. 132(5), pp. 521-528.

J. Ariffin. 2004. Development of sediment transport models for rivers in Malaysia using regression analysis and artificial neural networks. PhD Thesis, Universiti Sains Malaysia, Malaysia.

C.K. Chan, A. Ab. Ghani, N.A. Zakaria, Z. Abu Hasan and R. Abdullah. 2005. Sediment transport equation assessment for selected rivers in Malaysia. International Journal of River Basin Management. vol. 3(3), pp. 203-208.

O. Giustolisi, and D.A. Savic. 2006. A symbolic data driven technique based on Evolutionary Polynomial Regression. Journal of Hydroinformatics. vol. 8(3), pp. 207-222.

D.A. Savic, O. Giutolisi, L. Berardi, W. Shepherd, S. Djordjevic and A. Saul. 2006. Modelling sewer failure by evolutionary computing. Proceeding of the Institution of Civil Engineers, Water Management. pp. 111-118.

L. Berardi, O. Giustolisi, Z. Kapelan and D.A. Savic. 2008. Development of pipe deterioration models for water distribution systems using EPR. Journal of Hydro Informatics, vol. 10(2), pp. 113–126.

O. Giustolisi, A. Doglioni, D.A. Savic and F. Pierro. 2008. An evolutionary multiobjective strategy for the effective management of groundwater resources. Water Resources Research Journal. vol. 44, pp. 1-14.

N.R. Draper, and H. Smith. 1998. Applied regression analysis. New York: John Wiley and Sons.

J.R. Koza. 1992. Genetic programming: on the programming of computers by means of natural selection. London, England: A Bradford book, The MIT Press, Massachusetts.

A. Watson, I. Parmee. 1996. System identification using genetic programming, Proceedings of ACEDC'96, PEDC, University of Plymouth United Kingdom.

D. Laucelli, O. Giustolisi. 2011. Scour depth modelling by a multi-objective evolutionary paradigm. Engineering Modelling and Software. Vol. 26, pp. 498-509.

D.E Goldberg. 1989. Genetic algorithms in search, optimization and machine learning, Massachussets: Addison Wesley.

D. Laucelli, L. Berardi and A. Dogliono. 2009. Evolutionary polynomial regression (EPR) - toolbox, Version 2.0 SA (Stand alone version), Department of Civil and Environmental Engineering, Technical University of Bari, Italy.

M.A. Shahin, H.R. Maier and M.B. Jaksa. 2004. Data division for developing neural networks applied to geotechnical engineering. Journal of Computing in Civil Engineering, ASCE. vol. 18(2), pp. 105-114.

D.R. Legates and Jr. G.J.McCabe. 1999. Evaluating the use of "Goodness-of-Fit" measures in hydrologic and hydroclimatic model validation. Water Resources Research. vol. 35(1), pp. 233-241.

M.A.Shahin, H.R. Maier and M.B. Jaksa. 2004. Investigation into the robustness of artificial neural networks for a case study in civil engineering. International Congress on Modelling and Simulation (2005: Melbourne,Victoria), Modelling and Simulation Society of Australia and New Zealand.

A. Shaqlaih, L.White, and M. Zaman. 2011. Resilient modulus modeling with information theory approach. International Journal of Geomechanics,