MODELLING STREAMFLOW-SEDIMENT RELATIONSHIP USING GENETIC PROGRAMMING

ADESOJI TUNBOSUN JAIYEOLA

Department of Civil Engineering and Surveying

Mangosuthu University of Technology

Durban

SOUTH AFRICA

Corresponding author: jaiyeola@mut.ac.za, soj707@yahoo.com

Abstract: - The presence of sediment in a river or reservoir is detrimental to the operation and management of water resources because it affects the design, planning and management of any water resource. Hence it is important to accurately estimate the quantity of sediment flowing in a river or been transported into a reservoir. The process of measuring the quantity of sediment in a river manually or using automatic sampling device is labour intensive, expensive and time consuming. In this study a data- driven approach, genetic programming techniques is used to develop an explicit model that accurately captures the relationship between streamflow and suspended sediment. The accuracy of the developed models was evaluated using Root Mean Square Error (RMSE) and Determination Coefficient (R^2). The results show that GP is capable of modelling streamflow-sediment process accurately with R^2 value of 0.999 and RMS errors of 0.032 during the validation phase.

Key-Words: - Streamflow; suspended sediment; genetic programming; GPdotNET; data-driven; modelling.

1. Introduction

The functionality and management of water resources is greatly affected to a large extent by the presence of sediment in reservoirs. Among other things it affect hydroelectric-equipment longevity, reservoir filling, channel navigability, fish habitat and river aesthetics [1]. Hence it is very important to correctly predict the quantity of sediment flowing in a river and also been transported into a reservoir as this has a great impact on its design, planning and management [2]. The process of measuring the quantity of sediment in a river manually or using automatic sampling device in a monitoring network is very labour intensive, very expensive and also very time consuming [3]. Also Hydro-climatologic and hydrological forecasting has always been a challenge to hydrologist because of its dynamism, high complexity and non- stationarity nature. As a result there has been an increase in scientific approaches to predictive modelling, among these approaches is data-driven modelling [4]. This approach involves the development of simple mathematical equations which represent the relationship between variables from the analysis of their concurrent input and output time series [5]. There are several variables that influence the transportation of sediment into a reservoir, some of which are turbidity, precipitation, temperature, streamflow, rainfall, flow depth, particle density, hydraulic radius, sediment size, mean flow velocity, acceleration due to gravity, shear stress, kinematic

viscosity, density of water, volumetric concentration of sediment, cross-section geometry, bed roughness, friction factor with sediment, bed slope etc [6] and also there have several attempts to determine the relationship between the quantity of sediment transported by a river into a reservoir to those flow conditions. However the most acceptable relationship between these variable has been between stream flows and sediments [7]. The aim of this study is to use a data-driven modelling approach, genetic programming (GP), to determine the relationship between discharge and sediment in a reservoir and also to predict the quantity of sediment in a river flowing into a reservoir. GP has been widely use in solving many problems in engineering, science applications, artificial intelligence, mechanical and industrial models [7]. GP has been successfully applied and verified generally in the field of water resources engineering. The performance of GP, multi-linear regression, and conventional Sediment rating curve techniques to predict suspended sediment was investigated by [7] and the results shows that GP performed better. GP was also used by [8] to successfully predict the local scour downstream of hydraulic structures. An extension of GP, Gene Expression Programming (GP), was compared with Adaptive Neuro-Fuzzy Inference System (ANFIS) by [9] for the prediction of ground water table fluctuation and GEP predictions were more accurate. It was used by Zahra Zangeneh, Sirdaru [10] to investigate its ability as a new approach for estimation of bed load transported in Kuras River in Malaysia. The evolved model obtained high accuracy for both testing validation set and confirming its ability to successfully predict bed load transportation. Kisi and Shiri [1] used GP to estimate suspended sediment in a river, using daily flow and sediment load as input data, the result was compared with results from Artificial Neural Network ANN, Support Vector Machine (SVM) and Adaptive Neuro-Fuzzy Inference System (ANFIS) models and it was found that GP model was more accurate. Linear genetic programming, which is also an extension to GP, was used by Kisi and Guven [11] to estimate suspended sediment concentration carried by a river and it proved to be superior in estimating daily suspended sediment concentrations than the best neuro-fussy model. The use of GP for symbolic regression as an effective method for prediction and estimation in software engineering was investigated and compared with regression / machine learning models by Afzal and Torkar [12] and the study provided evidence in support of GP being an effective technique for software fault prediction and software reliability growth modelling. Genetic programming was also used by Ghorbani, AytekR [13] to forecast average sea water level values. Genetic programming and ANFIS models were used by Kisi and Shiri [14] to estimate river flow both for short and long term. So a number of applications of GP have been reported in water resources, which also includes include rainfall–runoff modelling [15, 16]; effect of flexible vegetation on flow in wetlands [17]; analysis and prediction of algal blooms [18, 19]. Its successful applications as an hydrological model can be found in [20-22]. Therefore GP can be effectively applied to the following areas, where (i) small enhancements in performance are easily and routinely measured (ii) analytical solutions are not provided by conventional mathematical analysis (iv) an estimated solution is acceptable and it's the best available (v) there are no clear understanding of the interrelationships between relevant variables, (vi) there is need to classification, integration and examine large amount of computer readable data, (vii) it is difficult to find the size and shape of the ultimate solution [23]. Genetic programming models have been found to be exceptionally good as regression tools especially for pattern recognition and complex non-linear estimations. It also reduces risk of over-fitting while training data so it is used in this research to estimate the quantity of suspended sediment flowing into Inanda Dam

2. Methodology

Genetic programming (GP) [24], which is derives from genetic algorithms is a systematic, domain-independent method that generates computer programs to solve problems automatically giving it a high level of what is expected from it [25]. GP involves a repeated random search for solution from an existing pool of computer programs which are potential solutions by applying the principle of natural evolution such as cross over and mutation to form a new population. This process continues until the best solution is obtained. These programs are expressed in form of a syntax tree where the nodes represent the instructions called the functions and the leaves which are the terminals represent the independent variables and random constants. Five preliminary steps are necessary before the operation of GP. They include the determination of (i) the terminal set; (ii) the functional set; (iii) the fitness measure; (iv) the parameters for controlling the run;

ISBN: 978-1-61804-338-2

and (v) the termination criterion and method of designating the result of the run [26]. The algorithm was also illustrated with a typical flow chart in figure 1 by [27]_ENREF_26.

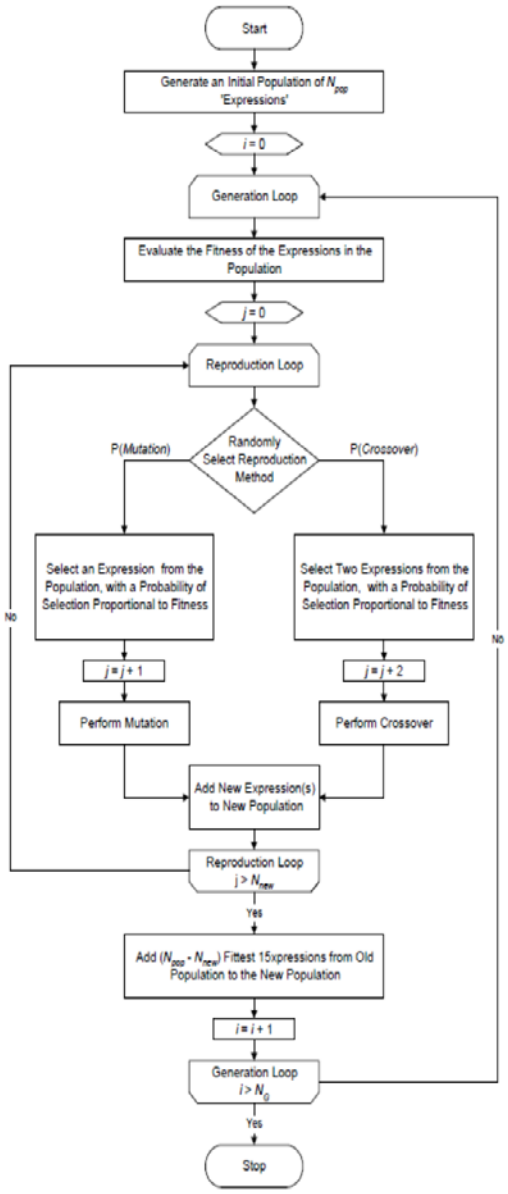

Fig.1: Flow sheet for Genetic programming. [27]

3. Case study

Inanda Dam was completed in 1989 and it is located within the valley of a thousand hills which is approximately 42km north of Durban. It is 23km long, 1.5km wide at its widest point and 50 meters deep. The water surface covers 1440 hectares. The Dam annual precipitation level is between 800mm - 1125mm and its temperatures range from 25°C to 38°C during summer and between from 9°C to 19°C during winter. The Dam is supplied with water by a large amount of perennial streams flowing from its surrounding but its main source of water is Umgeni River. The catchment area at this site is 4079 km^3 on latitude 29.70891 and longitude 30.86707. To develop the model for the estimation of suspended sediment in the Dam using GP, stream flow and suspended sediment dataset from 1983 to 2014 from the dam (RMG017-Mgeni/New Inanda weir) and its catchment area (station 02410729,Mount Edgecombe) were used in this study. These data were acquired from Umgeni Water (South Africa), South Africa weather Service (SAWS) and Department of Water Affairs (South Africa) website.

4. Application and results

In this study, five input variables were used in the GP input space. These include up-streamflow values for a given month and the last two months' streamflow values (Q_t, Q_{t-1}, Q_{t-2}) and the corresponding monthly suspended sediment values for the last two months (SS_{t-1}, SS_{t-2}). The target output is the suspended sediment value (SS_t) for the given month, where the subscript't' represents the given time period (month). Four input combinations of these five variables were used to develop the suspended sediment model for each month. For data splitting 75% of the whole data was used as the training set of the developed model while the remaining 25% of the whole data was used as its testing set. The accuracy of the developed models were evaluated using the Root Mean Square Error (RMSE), this is expressed mathematically as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(S_M - S_O)^2} \quad (1)$$

where S_O and S_M are the observed and predicted values of stream flow at time i, respectively. Root mean square error (RMSE) values ranges between 0and ∞, with lower values corresponding to better performance of the model. RMSE describes the average magnitude of the errors (differences between the observational values and model

results). The models were also evaluated using Determination Coefficient (R^2) expressed mathematically as :

$$R^2 = \left[\frac{\sum_{i=1}^{n}(S_0 - \bar{S}_0)\,(S_M - \bar{S}_M)}{\sqrt{\sum_{i=1}^{n}(S_0 - \bar{S}_0)^2 \sum_{i=0}^{n}(S_M - \bar{S}_M)^2}} \right]^2 \quad (2)$$

where So is the observed suspended sediment at the ith time step, S_M is the corresponding simulated suspended sediment, n is number of time steps, $\overline{S}_o$ is the mean of the observed values and $\overline{S}_M$ is the mean value for simulations. The Coefficient of Determination (R_2) values ranges between 0 and 1, with higher values indicating the better performance of the model.

GPdotNET programing software was used to run the simulations. The performance of the models were evaluated and from the results it can be established that all the models produced very low RMSE and very high R^2 values both during the training and validation phases demonstrating the accuracy of GP. The results also agrees with the results from the study conducted [28] confirming that GP models can predict accurately both during normal and extreme events. The models performed very well during the training phase with R^2 values of 0.994 to 0.999 and RMS errors of 0.026 between 5.998. This excellent perform was repeated during the validation phase with R^2 values of 0.996 to 0.999 and RMS errors from 0.032 to 2.382. This shows a strong and positive correlation between the measured and predicted suspended sediment during the training and validation phase as illustrated in figure 2. The observed and GP-predicted suspended sediment load (mg/l) during the validation phase is presented in Figure 2. This is the graphical and visual presentation of the output (suspended sediment) from the GPdotNET programme used in this study for validation phase

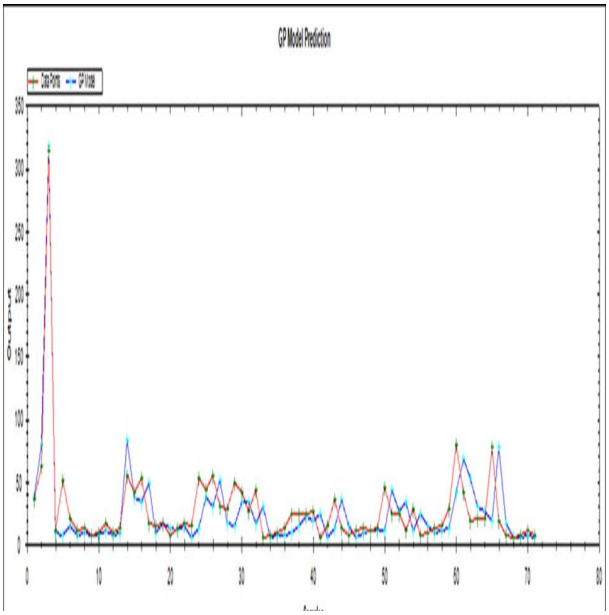

Fig. 2: The observed and estimated suspended sediments during the validation phase.

Figure 3 shows an expression tree of the developed GP model for the upstream station which is expressed as: (X1+((((Log10((Log10(R6))))+X1)/((R4+R4)*((R6/R5)-(R1+R4))))/(((Log10((R1/R3)))+R5)-((X1/R5)/(R6-R1))))) where R1= 1.75956,R2= 3.18723, R3= 1.36446, R4= 0.40293, R5= 9.33643, R6= 8.97514 and X1 =suspended sediment for time t.

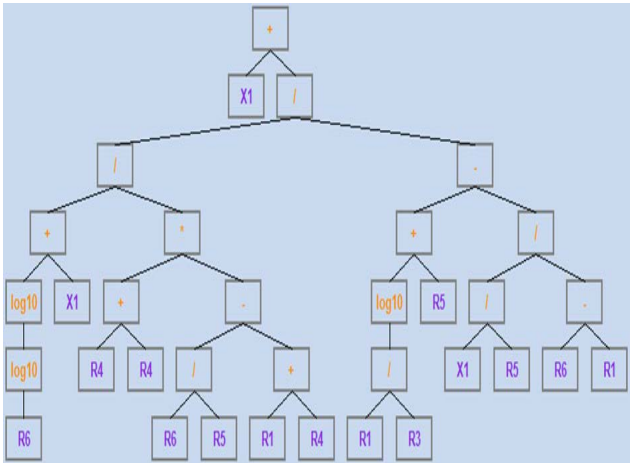

Fig. 3: Expression tree for the developed model.

5. Conclusion

The results show that the predictions of the GP models were very accurate especially in predicting large quantities of suspended sediment load during high streamflow such as during flood events. This proves that the use of GP is an accurate and superior alternative technique for the prediction of sediment load in a small or medium basin like Inanda dam. The results from this study also show the ability of the GP technique to capture the relationship between sediment load and streamflow in form of a simple and explicit model that can used by anyone. Therefore GP technique can be applied to real-time forecasting on short-basis like daily or hourly suspended sediment load predictions. The results above are part of an ongoing research work and it will be compared with those from other predictive tools. It is recommended that GP user should be trained to exploit it expanding capabilities and its integration with Geographical Information System (GIS) for Multi-model simulation is also highly advisable.

References

1. Kisi, O. and J. Shiri, *River suspended sediment estimation by climatic variables implication: Comparative study among soft computing techniques.* Computers & Geosciences, 2012. **43**: p. 73-82.
2. Kisi, O., et al., *Suspended sediment modeling using genetic programming and soft computing techniques.* Journal of Hydrology, 2012. **450-451**: p. 48-58.
3. Chang, N.-B. and Z. Xuan. *Monitoring nutrient concentrations in Tampa Bay with MODIS images and machine learning models*. in *Networking, Sensing and Control (ICNSC), 2013 10th IEEE International Conference on*. 2013. IEEE.
4. Nourani, V., et al., *Applications of hybrid wavelet–Artificial Intelligence models in hydrology: A review.* Journal of Hydrology, 2014. **514**(0): p. 358-377.
5. Solomatine, D. and A. Ostfeld, *Data-driven modelling: some past experiences and new approaches.* Journal of hydroinformatics, 2008. **10**(1): p. 3-22.
6. Ghani, A.A., *Sediment transport in sewers.* 1993.
7. Aytek, A. and O. Kisi, *A genetic programming approach to suspended sediment modelling.* Journal of Hydrology, 2008. **351**: p. 288– 298.
8. Guven, A. and M. Gunal, *Genetic programming approach for prediction of local scour downstream of hydraulic structures.* Journal of Irrigation and Drainage Engineering, 2008.
9. Shiri, J. and Ö. Kişi, *Comparison of genetic programming with neuro-fuzzy systems for predicting short-term water table depth fluctuations.* Computers & Geosciences, 2011. **37**(10): p. 1692-1701.
10. Zahra Zangeneh, Z., et al., *Sustainable Solution for Global Crisis of Flooding, Polluton and Water Scacity*, in *3rd International Conference on Managing Rivers in the 21st Century* 2007.
11. Kisi, O. and A. Guven, *A machine code-based genetic programming for suspended sediment concentration estimation.* Advances in Engineering Software 2010. **41**: p. 939–945.
12. Afzal , W. and R. Torkar, *On the application of genetic programming for software engineering predictive modelling.* Expert Systems with Applications, 2011. **38**: p. 11984–11997.
13. Ghorbani, M.A., et al., *Sea water level forecasting using genetic programming and comparing the performance with artificial neural networks.* Computers & Geosciences 2009.
14. Kisi, O. and J. Shiri. *A comparison of genetic programming and ANFIS in forecasting daily, monthly and daily streamflows*. in *Proceedings of the international symposium on innovations in intelligent systems and applications*. 2010.
15. Whigham, P.A. and P.F. Crapper, *Modelling rainfall–runoff relationships using genetic programming.* Mathematical and Computer Modelling 2001. **33 (6–7)**: p. 707–721.
16. Khu, S.T., et al., *Genetic programming and its application in real-time runoff forecasting.* Journal of American Water Resources Association 2001(37 (2)): p. 439–451.
17. Babovic, V. and M. Keijzer, *Declarative and preferential bias in GEP-based scientific discovery*. Genet Program Evolv 2002. **3(1)**: p. 41–79.
18. Muttil, N. and K.W. Chau, *Neural network and genetic programming for modelling*

coastal algal blooms. International Journal of Environment and Pollution, 2006. **28 (3–4)**: p. 223–238.
19. Muttil, N. and J.H.W. Lee *Genetic programming for analysis and real-time prediction of coastal algal blooms.* Ecological Modelling, 2005. **189 (3–4)**: p. 363–376.
20. Garg, V. and V. Jothiprakash, *Modeling the time variation of reservoir trap efficiency*. Journal of Hydrologic Engineering, 2010. **15**(12): p. 1001–1015.
21. Sivapragasam, C., R. Maheswaran, and V. Venkatesh, *"Genetic programming approach for flood routing in natural channels*. Hydrological Processes, 2008: p. 623–628,.
22. Parasuraman, K., A. Elshorbagy, and S.K. Carey, *Modelling the dynamics of the evapotranspiration process using genetic programming.* Hydrological Sciences Journal 2007. **52 (3)**: p. 563–578.
23. Banzhaf, W., P, R.E. Keller, and F.D. Francone, *Genetic programming: an introduction*. 1998, San Francisco (CA): Morgan Kaufmann.
24. Koza, J.R., *Genetic Programming: vol. 1, On the programming of computers by means of natural selection*. Vol. 1. 1992: MIT press.
25. Poli, R., et al., *A field guide to genetic programming*. 2008: Lulu. com.
26. Burke, E.K. and G. Kendall, *Search methodologies: introductory tutorials in optimization and decision support techniques*. 2005: Springer.
27. Willis, M.J., et al. *Genetic programming: An introduction and survey of applications*. in *IEE Conference Publications*. 1997. [London, England]: IEE, 1964-.
28. Londhe, S. and S. Charhate, *Comparison of data-driven modelling techniques for river flow forecasting.* Hydrological Sciences Journal–Journal des Sciences Hydrologiques, 2010. **55**(7): p. 1163-1174.