Journal of Civil Engineering and Urbanism

Volume 3, Issue 3: 104-106 (2013) (Received: January 13, 2013; Accepted: April 15, 2013; Published: May 30, 2013) ISSN-2252-0430

Evaluation of Genetic Programming for modeling solute breakthrough curve through the temporal data assignment scenarios

Sepideh Karimi[1]*, Ali Ashraf Sdaraddini[2], Amir Hossein Nazemi[2], Reza Delear Hasannia[2], Ozgur Kisi[3]

1M.Sc student of water Irrigation and Drainage, University of Tabriz, Tabriz, Iran
[2]Water Engineering Department, Faculty of Agriculture, University of Tabriz, Tabriz, Iran
[3]Civil Engineering Department, Architecture and Engineering Faculty, Canik Basari University, Samsun, Turkey

*Corresponding Author's E-mail address: karimi_sepide@yahoo.com

ABSTRACT: A modeling procedure was assessed in the present paper to investigate the abilities of Gene Expression Programming (GEP) approach for modeling solute breakthrough curve. The evaluation of the GEP method for modeling solute breakthrough curve was carried out through complete data scanning techniques. In this way, a complete scan of the possible train and test set configurations was carried out according to temporal criteria using 'leave one out' procedures. The obtained results reveal that the suitable assessment of the model performance should consider a complete temporal and/or spatial scan of the data set used.

Keywords: Breakthrough Curve, Bromide, Gene Expression Programming, Leave One Out

ORIGINAL ARTICLE

INTRODUCTION

Effective groundwater management is crucial from many aspects viewpoints as it makes decisive rules in agricultural systems management and planning, urban planning, drinking water withdrawn and management and study of the effect of industrial/domestic pollutions in groundwater pollution. Nevertheless, characterization of the contaminant transport through saturated/unsaturated soil layers is of primary importance in groundwater management. So far, various physical based models have been developed for modeling contaminant transport modeling using convection-dispersion equation. However, the validity of these models is limited for the cases the contaminant mixing is through the soil column in which the velocities are non uniform (Yoon et al., 2007).

In the recent years, application of Artificial Intelligence (AI) approaches [e.g. Genetic programming (GP)] has become viable in wide aspects of water resources systems forecasting and management.

GEP (Gene Expression Programming) is comparable to GP but involves computer programs of different sizes and shapes encoded in linear chromosomes of fixed lengths. There are two main players in GEP (Ferreira, 2006a): chromosomes (which are usually composed of more than one gene of equal length) and expression trees (programs) which are expressions of the genetic information encoded in chromosomes. The chromosomes are composed of multiple genes, each gene encoding a smaller subprogram. Furthermore, the structural and functional organization of linear chromosomes allows for unconstrained operation of important genetic operators, such as mutation, transposition and recombination. One strength of the GEP approach is that its search operators can always generate a valid structure and are suited to genetic diversity (Ferreira, 2006b). Another strength of GEP consists in its unique multigenic nature, which allows for the evolution of more complex programs composed of several subprograms. As a result, GEP surpasses the old GP system by 100-10,000 times (Ferreira, 2001). The most important advantages of GEP are (Ferreira, 2001b): (i) the chromosomes are simple entities: linear, compact, relatively small, easy to manipulate genetically (replicate, mutate, recombine, etc.); (ii) the expression trees are exclusively the expression of their respective chromosomes; they are entities upon which selection acts, and according to fitness, they are selected to reproduce with modification. According to Ferreira (Ferreira, 2006a) GEP is like GAs and GP, a genetic algorithm as it uses populations of individuals, selects them according to fitness, and introduces genetic variation using one or more genetic operators.

Various aspects of GP applications in engineering issues have been reported in the literature including rainfall-runoff modeling (Kisi et al., 2013), predicting groundwater table depth fluctuations (Shiri and Kisi, 2011a) estimating daily pan evaporation (Shiri and Kisi, 2011b), precipitation forecasting (Kisi and Shiri, 2011), modeling river suspended sediment load (Kisi and Shiri, 2012), and predicting daily lake level variations (Kisi, Shiri, and Nikoofar, 2012).

Available GEP applications consider a single data set assignment when training and test sets are defined. The present paper aims at evaluating GEP technique for modeling breakthrough curve through a complete temporal data scanning. This is the first time application

To cite this paper: Karimi, S., Sdaraddini, A.A., Nazemi,A.H., Delear Hasannia,R., Kisi, O.2013. Evaluation of Genetic Programming for modeling solute breakthrough curve through the temporal data assignment scenarios. *J. Civil Eng. Urban.,* 3(3): 104-106.
Journal homepage http://www.ojceu.ir/main/

of GP (i.e. GEP) in literature for modeling solute breakthrough curve.

MATERIALS AND METHODS

Experimental set and used data

Four 40cm vertical soil columns comprising two clay samples and two sand samples were considered in the present study, in which one disturbed and one undisturbed soil column was tested per each soil sample. The top boundary of the domain was established as a constant flux boundary (a constant flux of $CaBr_2$ of 0.01 molar) and the bottom was established as a seepage boundary.

Gene Expression Programming

The procedure starts with a random generation of chromosomes of a certain program (initial population). Then the generated chromosomes are expressed and the fitness of each individual program is evaluated against a set of fitness cases (Ferreira, 2001a). The programs are then selected according to their own fitness (their performance in that particular environment). The process is repeated until a good solution is found for the phenomenon under study. In the present work the GeneXpro program was used to apply GEP.

The procedure to model breakthrough curve involves the next general step. The first set of investigations used in the GEP model is the selection of an appropriate fitness function which may be variously defined (as absolute error, relative error and correlation coefficient). The second step consists of choosing the set of terminals T and the set of functions F to create the chromosomes. In the current problem, the terminal set includes flow and concentration values. The choice of an appropriate function depends on the user's viewpoint. The function set $\left\{+,-,\times,\div,\sqrt[3]{},\sqrt{},\ln,e^x,x^2,x^3,\sin x,\cos x,Arctgx\right\}$ gives the best results among other function sets as discussed by Shiri et al. (Shiri et al., 2012). So the GEP models in the present study were established using this function set. The next step is to choose the chromosomal architecture. The commonly used values for this architecture are (Ferreira, 2001a): length of head, h=8, and three genes per chromosome. The fourth step is to choose the linking function which should be chosen as "addition" or "multiplication" for algebraic sub-trees (Ferreira, 2006a), but recent investigations have shown that the addition linking function gives optimal results when it is applied for linking the parse trees [e.g., 10, 12]. The final step is to choose the genetic operators which can be taken as the default values of GeneXpro program (Shiri and Kisi , 2011b). The parameters used per run are summarized in Table 3.

Leave one out procedures

Normally, a single data set assignment is considered for training and test sets when assessing the performance of an AI model. Nevertheless, conclusions drawn up from this approach might be misleading [13-14]. In this work, the GEP model was evaluated by considering temporal (TLOO) leave one data set out approaches [15-17]. Accordingly, TLOO was carried out independently per soil column. Therefore, first, a minimum temporary test period was defined as one observational point. Second, according to that test period, a temporary leave one out approach was applied, leaving at each stage a different point for testing until a complete scan of the series of that soil sample was fulfilled. This process was repeated for each experimental soil column.

RESULTS AND DISCUSSIONS

Figure 1 represents the *RMSE* indicator of each LOO approach for each soil sample. As could be foreshadowed, the estimations of those models relying on undisturbed

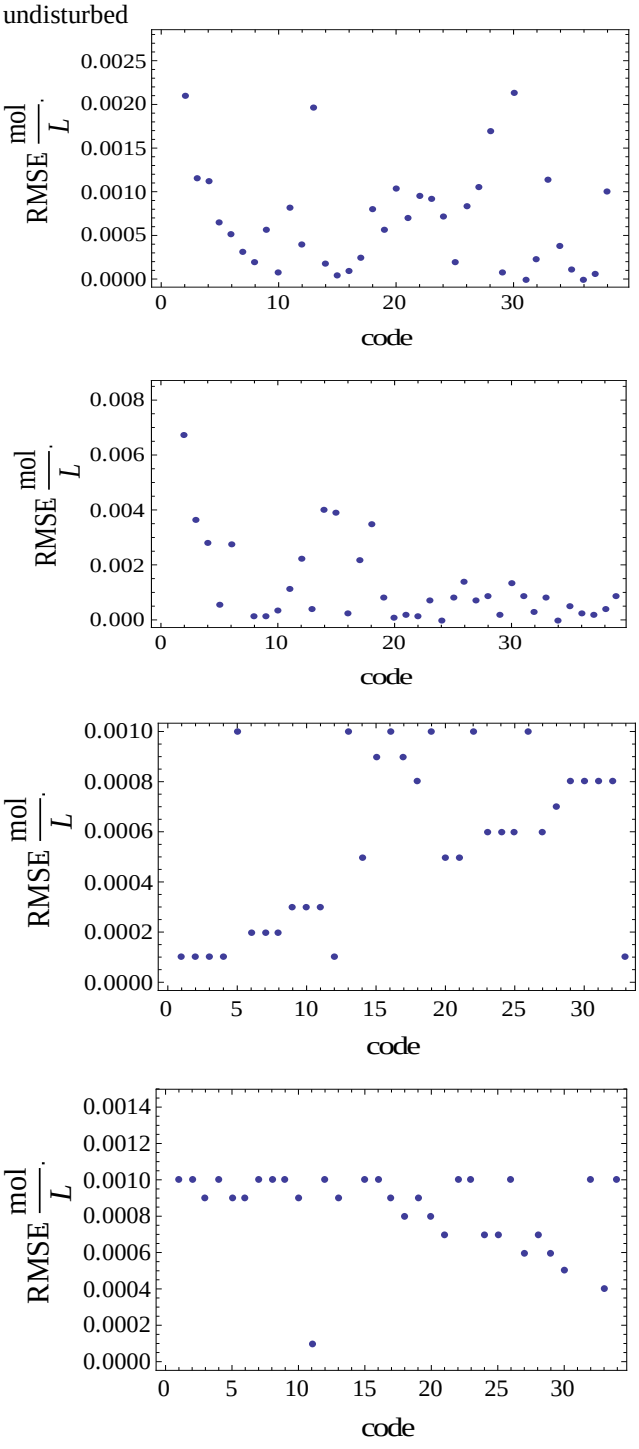

Figure 1. the performance indicator split up per test stage

To cite this paper: Karimi, S., Sdaraddini, A.A., Nazemi,A.H., Delear Hasannia,R., Kisi, O.2013. Evaluation of Genetic Programming for modeling solute breakthrough curve through the temporal data assignment scenarios. *J. Civil Eng. Urban.,* 3(3): 104-106.
Journal homepage http://www.ojceu.ir/main/

soil samples (i.e. clay-loam and sand-loam) present higher accuracy. Further, the LOO-GEP approach provides always the most accurate predictions, as it is trained with series of the same data used for testing (using different patterns for training and for testing). On the other hand, its generalization ability will be limited to similar conditions to those of the training (and testing) station.

Next, Figure 1 provides a pictorial representation of the performance indicator split up per test stage considering. In general, it can be stated, with some exceptions, that the individual accuracy of the disturbed soil samples models per test stage is considerably more fluctuating than for the undisturbed samples.

As mentioned above, it might be more suitable to consider LOO-GEP approaches, when there is a lack of observed variables in the train-test stages, being able to provide a model with high generalization ability. The performance fluctuations found out among soil samples highlight the need to assess the models performance through data set scanning procedures and not only considering a single data set assignment. Otherwise, the conclusions drawn up might be misleading.

Although, in this case, the order of accuracy of the models is qualitatively the same for all samples based on disturbance, there are important quantitative differences in the Δ*RMSE* ranges between approaches depending on the selected test stage. Hence, a complete testing scan of the data set is required to properly assess the performance of the estimations. Otherwise, what is a common practice, the conclusions can only be referred to the single test set assigned, which can be only partially valid.

Due to a higher input-output mapping ability, the GEP models are found to be more accurate. The difference in accuracy between the GEP models is lower for various soil samples. Thus, the GEP models improve when more inputs are considered. It is important to acknowledge the mapping ability of the SGEP models, because they provide sufficiently accurate estimates, even though they are trained without considering patterns of the test stage. Hence, suitably fed the GEP algorithms are able to acquire knowledge from training data and use it satisfactorily for estimation elsewhere.

REFERENCES

1. Yoon, H., Hyun, Y., Lee, K.K. 2007. Forecasting solute breakthrough curves through the unsaturated zone using artificial neural networks. Journal of Hydrology. 335: 68-77.
2. Ferreira, C. 2006a. Gene expression programming: Mathematical Modeling by an artificial intelligence. Springer, Berling, Heidelberg New York, 478 pp.
3. Ferreira, C. 2006b. Automatically defined functions in gene expression programming. In: Nedjah, N., Mourelleh, L.de M., Abraham, A. (eds)., Genetic Systems Programming: Theory and Experiences, Studies in Computational Intelligence, 13, 21-56, Springer, Verlag.
4. Ferreira, C. 2001a. Gene expression programming in problem solving. In: 6th Online World Conference on Soft computing in Industrial Applications (invited tutorial).
5. Ferreira, C. 2001b. Gene expression programming: a new adaptive algorithm for solving problems. Complex Systems 13(2): 87-129.
6. Kisi, O., Shiri, J., Tombul, M. 2013. Modeling rainfall-runoff process using soft computing techniques. Computers and Geosciences. 51: 108-117.
7. Shiri, J., Kisi, O. 2011a. Comparison of genetic programming with neuro-fuzzy systems for predicting short-term water table depth fluctuations. Computers and Geosciences 37(10): 1692-1701.
8. Shiri, J., Kisi, O. 2011b. Application of artificial intelligence to estimate daily pan evaporation using available and estimated climatic data in the Khozestan Province (Southwestern Iran). ASCE Journal of Irrigation and Drainage Engineering 137(7): 412-425.
9. Kisi, O., Shiri, J. 2011. Precipitation forecasting using wavelet-genetic programming and wavelet-neuro-fuzzy conjunction models. Water Resources Management 25(13): 3135-3152.
10. Kisi, O., Shiri, J. 2012. River suspended sediment estimation by climatic variables implication: comparative study among soft computing techniques. Computers and Geoscienes 43: 73-82.
11. Kisi, O., Shiri, J., Nikoofar, B. 2012. Forecasting daily lake levels using artificial intelligence approaches. Computers and Geosciences 41: 169-180.
12. Shiri, J., Kisi, O., Landeras, G., Lopez, J.J., Nazemi, A.H., Stuyt, L.C.P.M. 2012. Daily reference evapotranspiration modeling by using genetic programming approach in the Basque Country (Northwestern Spain). Journal of Hydrology 414-415: 302-316.
13. Marti, P., Manzano, J., Royuela, A. 2011. Assessment of 4-input artificial neural network for ET_0 estimation through data set scanning procedures. Irrigation Science 29: 181-195.
14. Shiri, J., Marti, P., Singh, V.P., 2013. Evaluation of gene expression programming approaches for estimating daily pan evaporation through spatial and temporal data scanning. Hydrological Processes, doi: 10.1002/hyp.9669.
15. Stone, M. 1974. Cross-validatory choice and assessment of statistical predictions. J Roy Statist Soc Ser B 36: 111-147.
16. Shao, J. 1993. Linear model selection by cross-validation, J Amer Statist Assoc 88(422):486-494
17. Hrachowitz, M., Soulsby, C., Imholt, C., Malcolm, I.A., Tetzlaff, D. 2010. Thermal regimes in a large upland salmon river: a simple model to identify the influence of landscape controls and climate change on maximum temperatures. Hydrological Processes 24: 3374-339

To cite this paper: Karimi, S., Sdaraddini, A.A., Nazemi,A.H., Delear Hasannia,R., Kisi, O.2013. Evaluation of Genetic Programming for modeling solute breakthrough curve through the temporal data assignment scenarios. *J. Civil Eng. Urban.,* 3(3): 104-106.
Journal homepage http://www.ojceu.ir/main/