Neural Comput & Applic (2017) 28 (Suppl 1):S207–S216
DOI 10.1007/s00521-016-2320-x

ORIGINAL ARTICLE

Estimation of soil dispersivity using soft computing approaches

Samad Emamgholizadeh[1] · Kiana Bahman[1] · S. Mohyeddin Bateni[2] · Hadi Ghorbani[1] · Isa Marofpoor[3] · Jeffrey R. Nielson[2]

Received: 15 October 2015 / Accepted: 6 April 2016 / Published online: 5 May 2016

Abstract The accurate estimation of soil dispersivity (α) is required for characterizing the transport of contaminants in soil. The in situ measurement of α is costly and time-consuming. Hence, in this study, three soft computing methods, namely adaptive neuro-fuzzy inference system (ANFIS), artificial neural network (ANN), and gene expression programming (GEP), are used to estimate α from more readily measurable physical soil variables, including travel distance from source of pollutant (L), mean grain size (D_{50}), soil bulk density (ρ_b), and contaminant velocity (V_c). Based on three statistical metrics [i.e., mean absolute error, root-mean-square error (RMSE), and coefficient of determination (R^2)], it is found that all approaches (ANN, ANFIS, and GEP) can accurately estimate α. Results also show that the ANN model (with RMSE = 0.00050 m and $R^2 = 0.977$) performs better than the ANFIS model (with RMSE = 0.00062 m and $R^2 = 0.956$), and the estimates from GEP are almost as accurate as those from ANFIS. The performance of ANN, ANFIS, and GEP models is also compared with the traditional multiple linear regression (MLR) method. The comparison indicates that all of the soft computing methods outperform the MLR model. Finally, the sensitivity analysis shows that the travel distance from source of pollution (L) and bulk density (ρ_b) have, respectively, the most and the least effect on the soil dispersivity.

Keywords Soil dispersivity · Adaptive neuro-fuzzy inference system · Artificial neural network · Genetic expression programming · Multiple linear regression

1 Introduction

Soil pollution or soil contamination occurs when harmful objects, chemicals, or substances are introduced directly or indirectly into the soil by humans. Such pollutants typically include agricultural chemicals (excess application of pesticides, herbicides, or fertilizer), heavy metals, industrial activity, inappropriate disposal of waste, and seepage from a landfill.

These pollutants, especially fertilizers, pesticides, and domestic and industrial wastewater, can infiltrate through the soil and ultimately reach the groundwater and degrade its quality. Because of the negative effects on groundwater and other potentially hazardous impacts, the movement and fate of contaminants in soil have received significant attention. Today, water pollution and soil pollution are main concerns of numerous countries and these countries are attempting to find solutions to mitigate these problems. The main concern about soil contamination is the risk it poses to human health [1].

The movement of contaminants in soil is described by dispersion and diffusion processes [2]. In an isotropic soil, dispersion occurs in two directions, transverse and longitudinal, with respect to the main flow. Two physical parameters, namely the longitudinal and transverse dispersion coefficients (D_L and D_T), are commonly used to characterize pollutant dispersion in soil [2, 3]. As the

✉ Samad Emamgholizadeh
s_gholizadeh517@Shahroodut.ac.ir

[1] Department of Water and Soil Engineering, Shahrood University of Technology, Shahrood, Iran

[2] Department of Civil and Environmental Engineering and Water Resource Research Center, University of Hawaii at Manoa, Honolulu, HI, USA

[3] Department of Water Engineering, University of Kurdistan, Sanandaj, Iran

longitudinal transfer of pollutants is more important than transverse dispersion, many studies (e.g., [2, 4, 5]) have focused on the longitudinal dispersion of pollutants and have used the one-dimensional advection–dispersion equation (ADE), which is given by Fried and Combarnous [3]:

$$\frac{dc}{dt} = D_L \frac{d^2c}{dl^2} - V\frac{dc}{dl} \quad (1)$$

where c is the concentration of pollutant in the soil (kg/m^3) at time t and distance l from the pollutant source, and V is the mean pore-water velocity (m/s). For a relatively homogeneous soil medium, D_L is linearly related to the pore-water velocity via [6],

$$D_L = \alpha V + D^* \quad (2)$$

where α is the soil dispersivity (m) and D^* is the molecular diffusion coefficient (m^2/s). For pore-water velocities (V) larger than 10^{-5} (m/s), αV becomes the dominate term, and thus, D_L can be approximated by αV (i.e., $D_L = \alpha.V$) [7].

The application of artificial intelligence (AI) models, such as genetic expression programming (GEP), adaptive neuro-fuzzy inference system (ANFIS), and artificial neural networks (ANN), has recently attracted the attention of researchers in soil science [8–11].

The in situ measurement of soil dispersivity is expensive, time-consuming, and difficult [4]. Hence, in this study, three soft computing techniques, namely ANFIS, ANN, and GEP, are applied to estimate the soil dispersivity. These methods can identify the complex relationship between inputs and outputs without previous knowledge of the governing equations between them.

Several studies have reported successful applications of ANN, ANFIS, and GEP for predicting the longitudinal dispersion coefficient (D_L) in rivers and streams. For example, Rowinski [12] used ANN to estimate D_L in rivers and showed that it can outperform physically based models. Toprak [13, 14] showed the superiority of the ANN model compared to regression-based equations for estimating D_L in channels. Tayfour and Singh [15] and Piotrowski [16] used ANN to estimate D_L in rivers. They compared the D_L estimates from ANN with those from the linear regression (LR) method and found better performance of the ANN model. Madvar [17] compared performance of the ANFIS model with empirical equations for forecasting D_L in natural streams and found that the ANFIS yields better results. Noori [18] applied a support vector machine (SVM) and ANFIS model for predicting the longitudinal dispersion coefficient in natural streams and showed that they performed better than the regression techniques. Recently, Sattar [19] used the GEP approach to predict D_L in the transitional and turbulent pipe flow. Using GEP, a relationship was developed between D_L and various influential pipe and flow characteristics such as average flow velocity, Reynolds number, pipe diameter, and the pipe friction coefficient.

The studies above indicate that ANN and ANFIS outperform regression-based techniques. To the best of our knowledge, no other study has used the ANN and ANFIS to predict D_L in porous media. The aims of this study are to (1) apply the ANN and ANFIS to estimate the soil dispersivity, (2) compare the performance of the ANN and ANFIS to the GEP and multiple linear regression (MLR) methods, and (3) determine the relative importance of influential variables on the soil dispersivity.

2 Models and methods

2.1 Artificial neural network (ANN)

The ANN mimics the nervous system of a human brain. The human brain receives many input signals and processes them to produce the proper output response. The ANN extracts implicit information contained in data without prior knowledge of the problem and utilizes it to solve problems [20–23].

The ANN consists of different layers (the input and output layers and one or more hidden layers) (see Fig. 1). These layers have neurons, which are linked to the neurons in the subsequent layer, to transfer signals from one layer to another. The ANN receives information through the input layer's node. For construction of ANN model, a user must select the number of nodes in the input, output, and hidden layers. The number of nodes in the input and output layers should be set equal to the number of input and output variables, respectively. Also, choosing the proper numbers of hidden layers and nodes is important, because with too many hidden layers and nodes, the network may memorize the relationship between inputs and outputs, rather than learning it. Such a network could be trained satisfactorily,

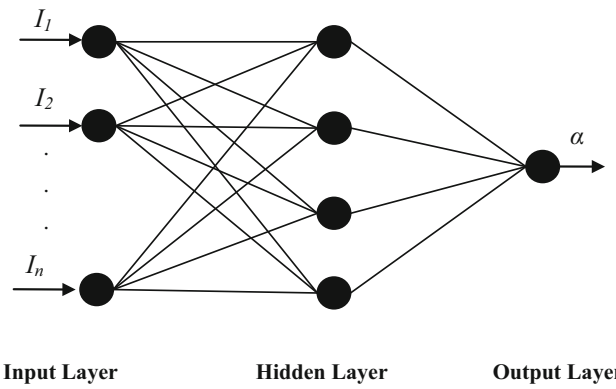

Fig. 1 Structure of a typical MLP model

but will likely perform poorly in the testing phase. Additionally, with too many hidden layers and nodes, the required training time increases significantly.

Among different kinds of ANNs (e.g., RBF, MLP, and Hopfield networks), the MLP is the most widely used feed-forward network in engineering problems [24–26]. In this study, we used the MLP with a backpropagation algorithm. The MLP network should be trained before application. The general structure of the MLP model with inputs I_1, I_2, …, and I_n and output α is shown in Fig. 1.

2.2 Adaptive neuro-fuzzy inference system (ANFIS)

The ANFIS is composed of the ANN and fuzzy system methods and was developed by Jang [27]. Based on the consequent of the fuzzy rules, the fuzzy inference systems are classified into three types: Tsukamoto's system, Sugeno's system, and Mamdani's system [27, 28]. The Sugeno fuzzy model is commonly used by researchers and thus is used in this study to predict the soil dispersivity.

In the Sugeno-type inference, two optimization methods, namely backpropagation and hybrid (composed of backpropagation and least squares), are used to update the membership function (i.e., Trimf, Gbellmf, Trapmf, Gaussmf, Pimf, Gauss2mf, Psigmf, and Dsigmf) [28, 29]. In the first-order Sugeno's system, if x and y are two inputs of the FIS, and f is its output, the two fuzzy IF/THEN rules can be shown as follows:

$$\text{Rule 1: If x is } A_1 \text{ and } y \text{ is } B_1 \quad \text{then } f_1 = p_1 x + q_1 y + r_1 \tag{3}$$

$$\text{Rule 2: If x is } A_2 \text{ and } y \text{ is } B_2 \quad \text{then } f_2 = p_2 x + q_2 y + r_2 \tag{4}$$

where A_1, A_2, B_1 and B_2 are the membership functions for inputs x and y,, respectively, and p_1, p_2, q_1, q_2, r_1, and r_2 are the parameters of the output function.

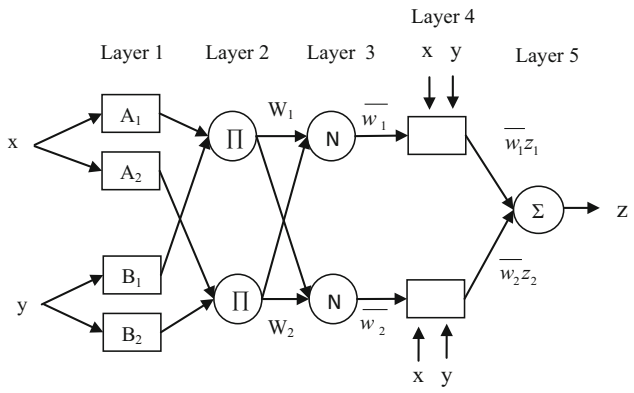

Fig. 2 Structure of ANFIS

Structure of the ANFIS model is shown in Fig. 2. For more details on the function of each node in any of the layers, interested readers can refer to the studies of Jang [27] and Madvar [17].

2.3 Gene expression programming (GEP)

GEP is an extension of genetic programming (GP) and genetic algorithm (GA) and was invented by Ferreria in 1999. It takes populations of individuals, chooses them based on their fitness, and uses one or more genetic operators to introduce genetic variations [30–33].

In recent years, GEP, GA, and GP models have been used widely in data mining applications. The GEP model has been shown to be superior to the GA and GP models. The advantages of the GEP model over the GA and GP models are inherent in its construction.

The first step to develop a GEP model is to select the fitness function, which can be created based on the error between model estimates and observations. The next step is to select a set of terminals, T (the variables or constants used in a problem), and functions, F [addition (+), subtraction (−), division (/), and multiplication (×)], to create the chromosomes. The third step is to design the chromosomal architecture, which includes picking the number of genes, the head size, and the linking function. The fourth step is to select the linking function for sub-ETs [32]. The fifth step is to implement the genetic operators (mutation, transposition, inversion, crossover/recombination, and gene crossover) for modification of the chromosomes. For more details on these operators, refer to Ferreira [32–34].

3 Data

3.1 Experiments

The soil dispersivity (α) depends on the travel distance from pollutant source (L), soil bulk density (ρ_b), porosity (n), hydraulic conductivity (K), mean grain size (D_{50}), and contaminant velocity (V_c) [35–39]. Experiments, 125 in total, were carried out in a rectangular Plexiglas tank [1.55 m (L) × 0.1 m (W) × 0.6 m (H)] to measure α for different L, ρ_b, n, K, D_{50}, and V_c values. Homogeneous sandy soils with five grain sizes (D_{50}) of 0.40, 0.60, 0.85, 1.20, and 1.70 mm were used in the experiments to represent a wide range of grain sizes. In each test, only one grain size was utilized. Sodium chloride (Nacl), with concentration of 9 g/l, was injected continuously as a nonreactive contaminant at velocities (V_c) of 7.2×10^{-5}, 9×10^{-5}, 11×10^{-5}, 14×10^{-5}, and 17×10^{-5} m/s. The samplings were taken at travel distances (L) of 0.25, 0.50, 0.75, 1, and 1.25 m from the injection point. Other

Table 1 Statistical indices of experimental data

Variable	Unit	Min	Max	Mean	SD	CV (%)
L	M	0.2500	1.2500	0.7500	0.3549	16.801
D_{50}	M	0.0003	0.0017	0.0009	0.0004	0.0219
ρ_b	kg/m^3	1590.0	1640.0	1618.0	16.064	15.949
n	–	0.3800	0.4000	0.3900	0.0063	0.0103
K	m/s	0.0009	0.0097	0.0051	0.0032	0.2124
V_c	m/s	0.0001	0.0004	0.0002	7.3158E−05	0.0024
α	M	0.0031	0.0168	0.0083	0.0029	0.1032

Table 2 The correlation coefficient (R) between the soil dispersivity (α) and its influential variables

	L	D_{50}	ρ_b	n	K	V_c	α
L	1						
D_{50}	0.1	1					
ρ_b	0.001	−0.658**	1				
n	−0.013	0.579**	−0.988**	1			
K	0.114	0.981**	−0.515**	0.431**	1		
V_c	−0.054	0.395**	−0.248*	0.203*	0.384**	1	
α	0.464**	0.895**	−0.509**	0.429**	0.889**	0.317**	1

** The relation is significant at the level 1 %

* The relation is significant at the level 5 %

influential variables, including soil bulk density (ρ_b), porosity (n), and hydraulic conductivity (K), were measured in each experiment. The minimum, maximum, mean, standard deviation (SD), and coefficient of variation (CV) of the experimental data are reported in Table 1.

3.2 Correlation analysis

Selection of appropriate input variables is important, because it affects the structure and performance of ANN, ANFIS, GEP, and MLR models. A correlation analysis was, therefore, carried out to evaluate correlation between the soil dispersivity (α) and the relevant soil variables (i.e., L, ρ_b, n, K, D_{50}, and V_c). As shown in Table 2, the strongest correlation was found between α and D_{50} with correlation (R) of 0.895. This finding is consistent with the studies of Xu and Eckstein [40] and Alipour and Kamanbedast [39].

A significant correlation ($R = 0.889$) was also observed between α and the hydraulic conductivity (K), which is consistent with the findings of Gelhar and Axness [41] and Xu and Eckstein [40]. The correlation analysis indicates that α is correlated with L, ρ_b, n, and V_c with an R of 0.464, −0.509, 0.429, and 0.317, respectively (Table 2). These results are consistent with the studies of Brigham [35], Neuman [42], Liu [36], Reinsch and Grossman [37], Xu and Eckstein [40], and Bromly [38].

As indicated in the literature (see also Table 2), α is correlated with L, ρ_b, n, K, D_{50}, and V_c, so they are used as inputs for the ANN, ANFIS, MLR, and GEP models. However, these models may be subject to collinearity, because the influential variables (i.e., L, ρ_b, n, K, D_{50}, and V_c) may have intercorrelation. The diagnostic results show that there is a significant correlation between K and D_{50} (with $R = 0.981$), and between ρ_b and n (with $R = 0.988$). Due to the significant intercorrelation, K and n were eliminated from the set of inputs to the models. Overall, based on the correlation analysis and the literature [37–39, 42], four variables, namely L, D_{50}, ρ_b, and V_c, were used to estimate α.

The 125 experimental data points were divided randomly into two parts: 80 % of the data (i.e., 100 data points) were used for training and the remaining 20 % (i.e., 25 data points) were utilized for testing.

3.3 Statistical metrics

The results of the ANFIS, ANN, GEP, and MLR models are evaluated using three statistical metrics, namely mean absolute error (MAE), root-mean-square error (RMSE), and coefficient of determination (R^2). These metrics are as follows:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{n}|O_i - P_i| \tag{5}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(O_i - P_i)^2}{N}} \tag{6}$$

$$R^2 = \frac{\sum_{i=1}^{N}(O_i - \overline{O})(P_i - \overline{P})}{\sqrt{\sum_{i=1}^{N}(O_i - \overline{O})^2 \sum_{i=1}^{N}(P_i - \overline{P})^2}} \quad (7)$$

where O_i and P_i are the observed and predicted values for the ith data point, $\overline{P}$ and $\overline{O}$ are the mean values of P_i and O_i, respectively, and N is the number of data points.

4 Results and discussion

4.1 Soil dispersivity estimates from the ANN model

As mentioned in Sect. 2.1, the MLP network with a backpropagation algorithm was used to predict the soil dispersivity. In each layer, the number of layers and nodes determines the structure of the MLP/ANN. In this study, there were four nodes in the input layer, corresponding to L, D_{50}, ρ_b, and V_c, and one node in the output layer, representing α. Also, for the hidden layer, choosing the correct number of nodes is important, because too few may result in underfitting, while too many may result in overfitting, and can increase the network training time. Following Singh [43], Bateni [44], and Emamgholizadeh [45], the trial-and-error approach was used to find the optimal number of nodes in the hidden layer.

Also four transfer functions, namely sigmoid ($f(x) = \frac{1}{1+e^{-x}}$), hyperbolic secant ($f(x) = \tanh(x)$), Gaussian ($f(x) = e^{-x^2}$), and hyperbolic tangent ($f(x) = \mathrm{Sech}(x)$), were used. The training of the ANN model was stopped when the RMSE, between the network's outputs and observations, was <0.0001 m, or when the maximum of 100,000 iterations was reached. Results show that the ANN model with one hidden layer, four hidden nodes, and the Gaussian transfer function gave the best results. Figure 3 illustrates the RMSE and R^2 of soil dispersivity (α) estimates from ANN versus the number of iterations in the training phase. Training of the model was stopped at 25,000 iterations, where the RMSE reached 0.00037 m.

The soil dispersivity estimates from the ANN model versus measurements in Fig. 4 for the training and testing phases are shown. As shown, the soil dispersivity estimates are close to the observations and fall around the 45° line. The ANN produced accurate soil dispersivity estimates, with the RMSE of 0.00040 m and 0.00050 m, in the training and testing phases, respectively. These results demonstrate that the ANN can learn the complex relationship between inputs (L, D_{50}, ρ_b, and V_c) and output (α).

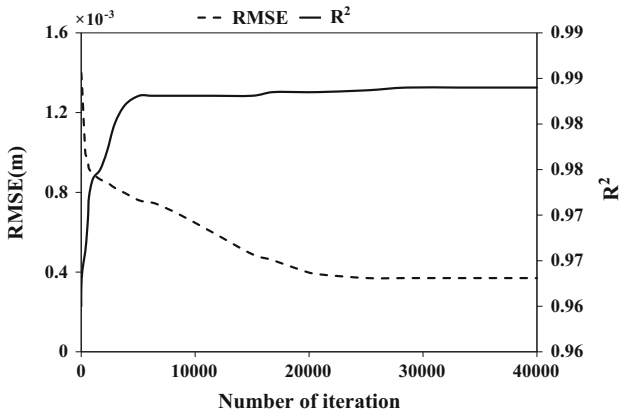

Fig. 3 Variation of RMSE and R^2 with iteration number in the training step

4.2 Soil dispersivity estimates from the ANFIS model

As mentioned in Sect. 2.2, there are three kinds of fuzzy inference systems, namely Mamdani, Sugeno, and Tsukamoto. In this study, the Sugeno-type with two optimization methods (i.e., backpropagation and hybrid) were used to update the membership functions [28]. Various types of membership functions (MFs) (i.e., Trapmf, Trimf, Gbellmf, Gaussmf, Gauss2mf, Dsigmf, Pimf, and Psigmf) were used to find the best structure for the ANFIS model. It was found

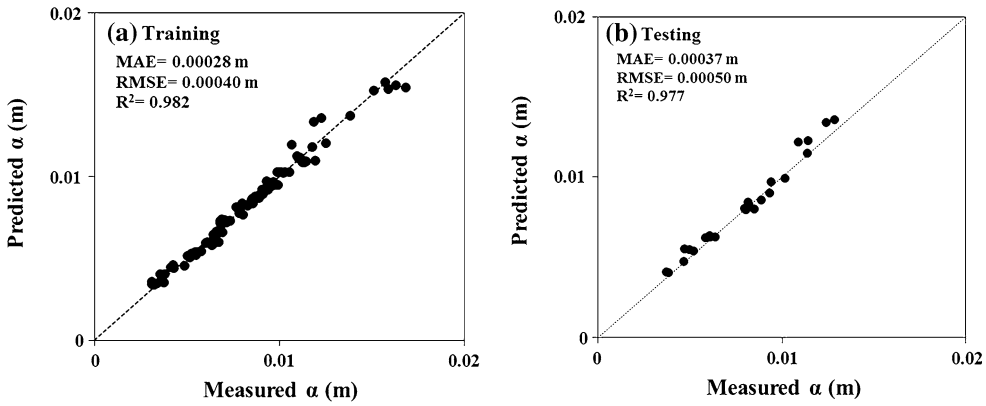

Fig. 4 Scatterplot of predicted versus measured soil dispersivity (α) from the ANN model

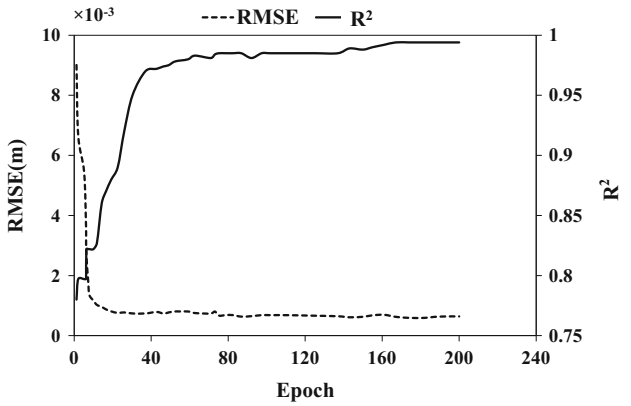

Fig. 5 Variation of RMSE and R^2 with iteration number in training step

that the Dsigmf input MF had the best results. Constant and linear functions were tried for the output MF. The results showed that the constant output MF performed better.

Figure 5 displays the RMSE and R^2 of soil dispersivity (α) estimates from ANFIS versus the number of iterations during the training step. Training of the ANFIS was stopped after 200 iterations, because changes in the RMSE were found to be negligible (see Fig. 5). After 200 iterations, the RMSE reached 0.00064 m and the variation in RMSE between two subsequent runs was 0.0001 m.

The trained ANFIS model was then used to predict soil dispersivity (α) values. The predictions were compared with the observations to evaluate the performance of the ANFIS model. Figure 6 compares the predicted α values with observations in the training and testing stages. As shown, the ANFIS model performed well in predicting soil dispersivity in both training (with MAE = 0.00048 m, RMSE = 0.00064 m, and $R^2 = 0.954$) and testing (with MAE = 0.00049 m, RMSE = 0.00062 m, and $R^2 = 0.956$) stages. As shown in Fig. 6, the ANFIS model is not as effective as the ANN, especially for predicting low and high α values. Overall, the estimates of soil dispersivity (α) from the ANFIS model are accurate, but they are not as good as those from the ANN model.

4.3 Soil dispersivity estimates from the GEP Model

The GEP model was used to estimate soil dispersivity. The first step was to choose the fitness function. In this study, the following fitness function (f_i) was used for the GEP model,

$$f_i = 1000 \times \frac{1}{1 + \mathrm{rMAE}_i} \tag{8}$$

where rMAE_i is the root mean absolute error of an individual chromosome i. f_i ranges from 0 to 1000 (1000 corresponds to a chromosome with ideal fitness) [34, 46]. Also, an initial population, which is composed of chromosomes, needs to be selected for the GEP model. In this study, multigenic chromosomes (including three genes) were used. Any population size can be used for the initial population, but studies (e.g., [31, 45–47]) have suggested using a value between 30 and 100. By trial and error, a population size of 50 chromosomes was selected as the optimum size.

After choosing a fitness function, functions and a set of terminals should be selected. Four basic arithmetic operators (+, −, ×, /) were used as functions, and the four input variables (i.e., L, D_{50}, ρ_b, and V_c) were utilized as terminals, $T = \{L, D_{50}, \rho_b, V_c\}$. Also, three genes per chromosome were used, and the head size was set to eight. The genetic operators containing inversion, mutation, transposition, recombination, and their rates were determined, after finalizing the chromosome architecture. The rate of the genetic operators is given in Table 3.

The last step was to select a linking function. It can be addition (+), multiplication (×), subtraction (−), or division (/) [31]. Here, the addition operator was selected as the

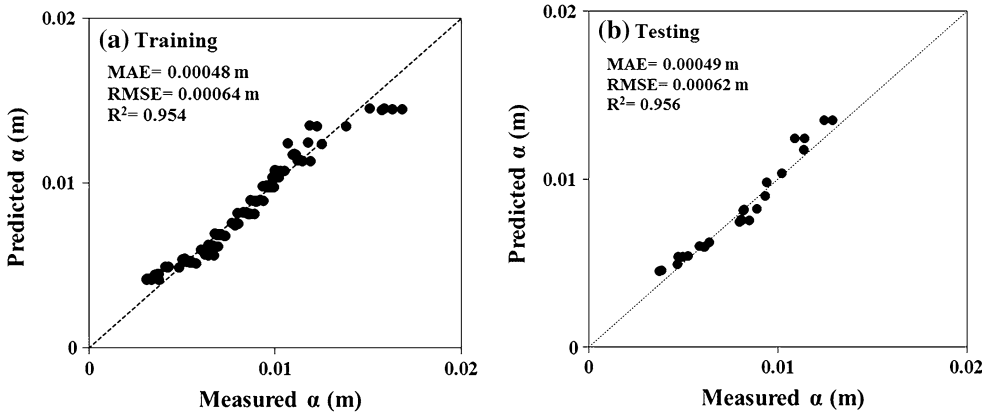

Fig. 6 Scatterplot of predicted versus measured soil dispersivity (α) from the ANFIS model

Table 3 Parameters of the optimized GEP model

Description of parameters	Setting of parameters
Mutation rate	0.044
Inversion rate	0.1
Insertion sequence (IS) transposition rate	0.1
Root insertion sequence RIS transposition rate	0.1
One-point and two-point recombination rate	0.3
Gene recombination rate	0.1
Gene transportation rate	0.1

linking function, due to its common use in other studies [48–52]. The iterations were stopped after 667,200 generations, because there was no increase in the fitness function value. The GEP-based equation for estimating soil dispersivity (α) is given by:

$$\alpha = [-1.941 + [(V_c^{1/3} - L)^2]^3] \times V_c + [(5.616\,V_c + L)^2]^{\frac{1}{3}} \times \text{Arctan}(D_{50}^{\frac{2}{3}}) + [\text{Arctan}\,(\text{Arctan}\,[\text{Arctan}\,(\cos [-0.131\rho_b/D_{50}])])] \times D_{50} \quad (9)$$

The form of Eq. (9) indicates that a nonlinear expression is needed for an accurate prediction of the soil dispersivity.

Figure 7 shows that the GEP makes accurate predictions of soil dispersivity, with MAE = 0.00035 m, RMSE = 0.00057 m, and $R^2 = 0.965$, in the training stage, and MAE = 0.00051 m, RMSE = 0.00066 m, and $R^2 = 0.964$, in the testing stage. This indicates that the GEP model outperformed the ANFIS model, in estimating soil dispersivity.

4.4 Comparing the soil dispersivity estimates from AI models with those of MLR model

To further assess the capability of ANFIS, ANN, and GEP models in estimating soil dispersivity (α), their results were compared with those of the MLR approach. MLR is used to create linear relations between dependent and independent variable (s). Here, the Statistical Package for the Social Sciences (SPSS) [53] was used to estimate the soil dispersivity. The MLR-based equation is given by,

$$\alpha = -0.028 + 0.003L + 5.986D_{50} + 0.177 \times 10^{-4}\rho_b - 0.141V_c \quad (10)$$

The MAE, RMSE, and R^2 of the soil dispersivity estimates from all approaches are presented in Table 4. As shown, the ANN, ANFIS, and GEP models outperformed MLR and had lower RMSE by 30.6, 13.9, and 8.3 %, respectively. The soil dispersivity estimates from ANN were better than those from the ANFIS and GEP models. Also, by a smaller margin, ANFIS outperformed GEP, as shown by a RMSE difference of 6.1 %.

Although the ANN and ANFIS models produced more accurate soil dispersivity estimates than GEP and MLR, they cannot generate an equation between α and the relevant soil properties. The ANN and ANFIS relate α to the relevant soil characteristics by a complex network that is

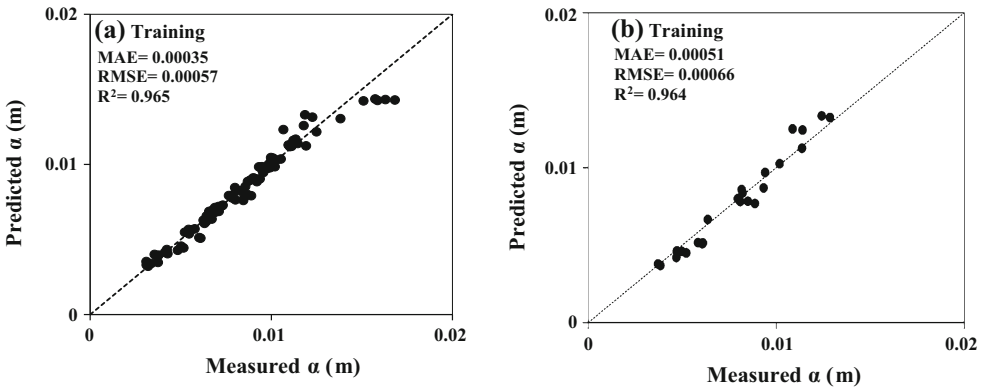

Fig. 7 Scatterplot of predicted versus measured soil dispersivity (α) from the GEP model

Table 4 Comparison of soil dispersivity estimates by ANN, ANFIS, GEP, and MLR

Model	Training			Testing		
	MAE × 10^{-3}	RMSE × 10^{-3}	R^2	MAE × 10^{-3}	RMSE × 10^{-3}	R^2
ANN	0.28	0.40	0.982	0.37	0.50	0.977
ANFIS	0.48	0.64	0.954	0.49	0.62	0.956
GEP	0.35	0.57	0.965	0.51	0.66	0.964
MLR	0.49	0.68	0.947	0.55	0.72	0.932

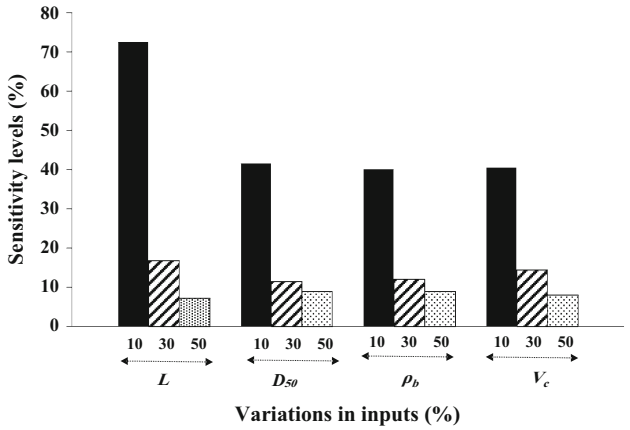

Fig. 8 Sensitivity levels of inputs on output

composed of transfer functions and weighting coefficients. In contrast, the GEP and MLR models can generate explicit expressions for estimation of α.

Overall, the results of this study indicate that the soft computing approaches (ANN, ANFIS, and GEP) are viable alternative techniques to the commonly used MLR model.

4.5 Sensitivity analysis

A number of sensitivity tests were performed to assess the relative importance of each input variable on soil dispersivity. Each input variable was varied by specified percentages (10, 30, and 50 %), and the corresponding changes in α were assessed. The sensitivity of soil dispersivity to changes in each input was quantified by using the following equation,

$$\text{Sensivity Level}\,(\%) = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{\%\,\text{change in output}}{\%\,\text{change in input}}\right) \times 100 \tag{11}$$

where $N = 25$ and is the number of testing data points. A higher sensitivity level implies a larger impact of input variable on the output. Figure 8 shows the sensitivity of α to input variables (i.e., L, D_{50}, ρ_b, and V_c).

As illustrated, α has the highest sensitivity to the travel distance from the pollutant source (L) and the lowest sensitivity to bulk density (ρ_b). By varying L by 10, 30, and 50 %, the α estimates change 72.63, 16.81, and 7.24 %, respectively. Similarly, by changing each input variable, the corresponding variation in the output was obtained. In summary, input parameter sensitivities from highest to lowest are listed as follows: the travel distance from the pollutant source (L), contaminant velocity (V_c), mean grain size (D_{50}), and bulk density (ρ_b).

5 Conclusion

Three soft computing techniques, namely ANFIS, ANN, and GEP, were employed to estimate soil dispersivity. The performance of these models was also compared to the MLR approach. A correlation analysis was performed to determine the correlation between soil dispersivity and relevant soil variables. Based on the results of the correlation analysis and existing studies (e.g., [37–39, 42]), four soil variables, namely travel distance from the pollutant source (L), mean grain size (D_{50}), bulk density (ρ_b), and contaminant velocity (V_c), were used as input variables.

A comparison of the results from the soft computing approaches showed that the ANN model outperformed the ANFIS and GEP models. Also, the accuracy of the GEP model to estimate soil dispersivity was slightly lower than the ANFIS model. It was found that soil dispersivity estimates from GEP were less accurate than those of ANN model. Next, the performance of the soft computing approaches was compared to the MLR model. The results show that the ANN, ANFIS, and GEP models estimated the soil dispersivity with the RMSE of 0.00050, 0.00062, and 0.00066 m, which is, respectively, 30.6, 13.9, and 8.3 % less than the RMSE of 0.00072 m from the MLR model. Finally, sensitivity analyses showed that the travel distance from the source of pollution (L) and bulk density (ρ_b) had, respectively, the greatest and the least effect on the soil dispersivity.

References

1. Dominguez JB (2008) Soil contamination research trends. Nova Publishers, New York. ISBN-13: 978-1604563191
2. Fried JJ, Combarnous MA (1971) Dispersion in porous media. In: Chow VT (ed) Advances in hydroscience. Academic Press, New York, pp 169–282
3. Bruggeman GA (1999) Analytical solutions of geohydro-logical problems. New York. Elsevier. ISBN 0-444-81829-4
4. Perfect E, Sukop MC, Haszler GR (2002) Prediction of dispersivity for undisturbed soil columns from water retention parameters. Soil Sci Soc Am J 66(3):696–701. doi:10.2136/sssaj2002.6960
5. Ujfaludi L (1986) Longitudinal dispersion tests in non-uniform porous media. Hydrol Sci J 31(4):467–474. doi:10.1080/02626668609491067
6. Freeze RA, Cherry JA (1979) Groundwater. Prentice-Hall. Englewood Cliffs, NJ. ISBN-13: 978-0133653120
7. Gillham RW, Cherry JA (1982) Contaminant migration in saturated unconsolidated geological deposits. Geol Soc Spec Pap 189:31–62. doi:10.1130/SPE189-p31GSA
8. Cal Y (1995) Soil classification by neural network. Adv Eng Softw 22(2):95–97. doi:10.1016/0965-9978(94)00035-H
9. Mukhlisin M, El-Shafie A, Taha MR (2012) Regularized versus non-regularized neural network model for prediction of saturated soil-water content on weathered granite soil formation. Neural Comput Appl 21(3):543–553. doi:10.1007/s00521-011-0545-2

10. Taghavifar H, Mardani A (2014) Use of artificial neural networks for estimation of agricultural wheel traction force in soil bin. Neural Comput Appl 24(6):1249–1258. doi:10.1007/s00521-013-1360-8
11. Yusof MF, Azamathulla HM, Abdullah R (2014) Prediction of soil erodibility factor for Peninsular Malaysia soil series using ANN. Neural Comput Appl 24(2):383–389. doi:10.1007/s00521-012-1236-3
12. Rowinski PM, Piotrowski A, Napiorkowski JJ (2005) Are artificial neural network techniques relevant for the estimation of longitudinal dispersion coefficient in rivers? Hydrol Sci J 50(1):175–187. doi:10.1623/hysj.50.1.175.56339
13. Toprak ZF, Cigizoglu HK (2008) Predicting longitudinal dispersion coefficient in natural streams by artificial intelligence methods. Hydrol Process 22(20):4106–4129. doi:10.1002/hyp.7012
14. Toprak ZF (2004) Determination of longitudinal dispersion coefficient in natural channel using fuzzy logic method. Ph.D. thesis, ITU, Institute of Science and Technology
15. Tayfur G, Singh VP (2005) Predicting longitudinal dispersion coefficient in natural streams by artificial neural network. J Hydraul Eng ASCE 131(11):991–1000. doi:10.1061/(ASCE)0733-9429(2005)131:11(991)
16. Piotrowski A (2005) Application of neural networks for longitudinal dispersion coefficient assessment. Geophys Res Abstr 7:00976 **(Hs1-1th5p-0003)**
17. Madvar HR, Ayyoubzadeh SA, Khadangi E, Ebadzadeh MM (2009) An expert system for predicting longitudinal coefficient in natural streams by using ANFIS. Expert Syst Appl 36(4):8589–8596. doi:10.1016/j.eswa.2008.10.043
18. Sattar A (2013) Gene expression models for the prediction of longitudinal dispersion coefficients in transitional and turbulent pipe flow. J Pipeline Syst Eng Pract. doi:10.1061/(ASCE)PS.1949-1204.0000153
19. Noori R, Karbassi A, Farokhnia A, Dehghani M (2009) Predicting the longitudinal dispersion coefficient using support vector machine and adaptive neuro-fuzzy inference system techniques. Environ Eng Sci 26(10):1503–1510. doi:10.1089/ees.2008.0360
20. Kashi H, Emamgholizadeh S, Ghorbani H (2014) Estimation of soil infiltration and cation exchange capacity based on multiple regression, ANN (RBF, MLP), and ANFIS models. Commun Soil Sci Plant 45(9):1195–1213. doi:10.1080/00103624.2013.874029
21. Haykin S (1994) Neural networks. A comprehensive foundation. IEEE press, MacMillan, New York. ISBN: 0023527617
22. Emamgholizadeh S, Parsaeian M, Baradaran M (2015) Seed yield prediction of sesame using artificial neural network. Eur J Agron 68:89–967. doi:10.1016/j.eja.2015.04.010
23. Azamathulla HM (2013) A review on application of soft computing methods in water resources engineering. Metaheuristics Water Geotech Transp Eng. doi:10.1016/B978-0-12-398296-4.00002-7
24. Rumelhart DE, Mcclelland JL, PDP research group (1986) Parallel recognition in modern computers. In: Proceeding: explorations in the microstructure of cognition. Foundations, MIT Press/Bradford Book, Cambridge Mass
25. Azamathulla HM, Deo MC, Deolalikar PB (2005) Neural networks for estimation of scour downstream of a ski-jump bucket. ASCE J Hydraul Eng ASCE 131(10):898–908. doi:10.1061/(ASCE)0733-9429(2005)131:10(898)
26. Jang JSR (1993) ANFIS-Adaptive-network-based fuzzy inference system. IEEE Trans Syst Sci Cybern 23(3):665–685. doi:10.1109/21.256541
27. Adewuyi PA (2012) Performance evaluation of Mamdani-type and Sugeno-type fuzzy inference system based controllers for computer fan. Int J Info Technol Comput Sci 5(1):26–36. doi:10.5815/ijitcs.2013.01.03
28. User's Guide of MATLAB (2002) Fuzzy logic toolbox for use with MATLAB. The MathWorks, Inc. Version 2.1.2, 1-244
29. Emamgholizadeh S, Kashi H, Marofpoor I, Zalaghi E (2014) Prediction of water quality parameters of Karoon River (Iran) by artificial intelligence-based models. Int J Environ Sci Technol (IJEST) 11(3):645–656. doi:10.1007/s13762-013-0378-x
30. Mitchell M (1996) An introduction to genetic algorithms. MIT press, London
31. Ferreira C (2001) Gene expression programming: a new adaptive algorithm for solving problems. Complex Syst J 13(2):87–129
32. Ferreira C (2001) Gene expression programming in problem solving. In: Invited tutorial of the 6th online world conference on soft computing in industrial applications, September 10–24
33. Zahiri R, Azamathulla HM, Ghorbani KH (2014) Prediction of local scour depth downstream of bed sills using soft computing models. Comput Intell Tech Earth Environ Sci. doi:10.1007/978-94-017-8642-3_11
34. Ferreira C (2006) Gene expression programming: mathematical modeling by an artificial intelligence, 2nd edn. Springer, Berlin. ISBN: 3540327967
35. Brigham WE (1974) Mixing equations in short laboratory columns. Soc Pet Eng J 14:91–99
36. Liu CCK, Loague K, Feng JS (1991) Fluid flow and solute transport processes in unsaturated heterogeneous soils: preliminary numerical experiments. J Contam Hydrol 7(3):261–283. doi:10.1016/0169-7722(91)90031-U
37. Reinsch TG, Grossman RB (1995) A method to predict bulk density of tilled Ap horizons. Soil Tillage Res 34(2):95–104. doi:10.1016/0167-1987(95)00458-5
38. Bromly M, Hinz C, Aylmore LAG (2007) Relation of dispersivity to properties of homogeneous saturated repacked soil columns. Eur J Soil Sci 58(1):293–301. doi:10.1111/j.1365-2389.2006.00839.x
39. Alipour R, Kamanbedast AA (2011) Investigation of vertical transmission of pollution at laboratory model and it's vitalizing for determination of dispersion coefficient at homogenous sandy soil. World Appl Sci J 14(2): 351–355. ISSN: 1818-4952
40. Xu M, Eckstein Y (1997) Statistical analysis of the relationships between dispersivity and other physical properties of porous media. Hydrogeol J 5(4):4–20. doi:10.1007/s100400050254
41. Gelhar LW, Axness CL (1983) Three dimensional stochastic analysis of macrodispersion in aquifers. Water Resour Res 19(1):161–180. doi:10.1029/WR019i001p00161
42. Neuman SP (1990) Universal scaling of hydraulic conductivities and dispersivities in geologic media. Water Resour Res 26(8):1749–1758. doi:10.1029/WR026i008p01749
43. Singh KP, Basant A, Malik A, Jain G (2009) Artificial neural network modeling of the river water quality, a case study. Ecol Model 220(6):888–895. doi:10.1016/j.ecolmodel.2009.01.004
44. Bateni SM, Borghei SM, Jeng DS (2007) Neural network and neuro-fuzzy assessments for scour depth around bridge piers. Eng Appl Artif Intell 20(3):401–414. doi:10.1016/j.engappai.2006.06.012
45. Emamgholizadeh S, Moslemi K, Karami G (2014) Prediction the groundwater level of bastam plain (Iran) by artificial neural network (ANN) and adaptive neuro-fuzzy inference system (ANFIS). Water Resour Manag 28(15):5433–5446. doi:10.1007/s11269-014-0810-0
46. Emamgholizadeh S, Bateni SM, Shahsavani D, Ashrafi T, Ghorbani H (2015) Estimation of soil cation exchange capacity using Genetic Expression Programming (GEP) and Multivariate Adaptive Regression Splines (MARS). J Hydrol. doi:10.1016/j.jhydrol.2015.08.025
47. Azamathulla HM, Ahmad Z (2012) Gene-expression programming for transverse mixing coefficient. J Hydrol 434–435:142–148. doi:10.1016/j.jhydrol.2012.02.018

48. Zahiri R, Dehghani AA, Azamathulla HM (2015) Application of gene-expression programming in hydraulic engineering. In: Gandomi AH et al (eds) Handbook of genetic programming applications. Springer International Publishing, Switzerland. doi:10.1007/978-3-319-20883-1_4
49. Guven A, Talu NE (2010) Gene expression programing for estimating suspended sediment yield in Middle Euphrates Basin, Turkey. Clean Soil Air Water 38(12):1159–1168. doi:10.1002/clen.201000003
50. Parhizkar S, Ajdari K, Kazemi GA, Emamgholizadeh S (2015) Predicting water level drawdown and assessment of land subsidence in Damghan aquifer by combining GMS and GEP models. Geopersia 5(1):63–80
51. Hashmi MZ, Shamseldin AY, Melville BW (2011) Statistical downscaling of watershed precipitation using gene expression programming (GEP). Environ Model Softw 26:1639–1646. doi:10.1016/j.envsoft.2011.07.007
52. Azamathulla HM, Jarrett RD (2013) Use of gene-expression programming to estimate manning's roughness coefficient for high gradient streams. Water Resour Manag 27(3):715–729. doi:10.1007/s11269-012-0211-1
53. SPSS Inc (2007) SPSS Base 16.0 applications guide. SPSS Inc., Chicago