Model. Earth Syst. Environ. (2016) 2:156
DOI 10.1007/s40808-016-0217-4

ORIGINAL ARTICLE

Determination of a suitable model for prediction of soil cation exchange capacity

Javad Seyedmohammadi[1] · Leila Esmaeelnejad[2] · Hassan Ramezanpour[3]

Received: 1 August 2016 / Accepted: 13 August 2016 / Published online: 20 August 2016

Abstract Analysis and design of land-use management scenarios requires detailed soil data. The cation exchange capacity (CEC) of soil is a basic chemical property, as it has been approved that the spatial distribution of CEC is important for decisions concerning pollution prevention, crop and farming management. Since laboratory procedures for measuring CEC are cumbersome and time-consuming, it is essential to develop an indirect approach such as pedotransfer functions to predict this parameter from more readily available soil data. The aim of this study was to compare multiple linear regression, multiple non-linear regression, adaptive neuro-fuzzy inference system and artificial neural network including feed-forward back propagation (FFBP) model to develop PTFs for predicting paddy soils CEC in Guilan province, northern Iran. Two soil parameters including organic carbon and clay were considered as input variables for proposed models. 171 soil samples were used. The data set was divided into two subsets for calibration and testing of the models. The models prediction capability was evaluated by comparison with observed data through various descriptive statistical indicators include root mean square error, determination coefficient, mean bias error and relative improvement values. Results showed that the FFBP model had the most reliable prediction when compared with other models and that provide a new methodology with acceptable accuracy to estimate the CEC of soil that diminished the engineering effort, time and funds and can provide the scientific basis for the study of soil CEC and be helpful for the estimation of soil CEC in other places with similar conditions, too.

Keywords Multiple linear regressions · Adaptive neuro-fuzzy inference system · Feed-forward back-propagation network · Cation exchange capacity

Abbreviations

ANFIS	Adaptive neuro-fuzzy inference system
ANN	Artificial neural network
CEC	Cation exchange capacity
CV	Coefficient of Variation
FFBP	Feed-forward back-propagation
FIS	Fuzzy inference system
MBE	Mean bias error
MF	Membership function
MLR	Multiple linear regressions
MNLR	Multiple non-linear regressions
MLP	Multi-layer perceptron
OC	Organic carbon
PTFs	Pedotransfer functions
R^2	Determination coefficient
RI	Relative improvement
RMSE	Root mean square error
SD	Standard deviation

Introduction

There is an increasing demand for reliable large-scale soil data to meet the requirements of models for planning of land-use systems, characterization of soil pollution, and

✉ Javad Seyedmohammadi
seyedmohammadi.javad@gmail.com

[1] Department of Soil Science, Faculty of Agriculture, University of Tabriz, Tabriz, Iran

[2] Department of Soil Science, Faculty of Agricultural Engineering and Technology, College of Agriculture and Natural Resources, University of Tehran, Karaj, Iran

[3] Department of Soil Science, Faculty of Agriculture, University of Guilan, Rasht, Iran

prediction of land degradation (McBratney et al. 2002; Zolfaghari et al. 2016). Cation exchange capacity (CEC) is one of the most important soil properties that is required in soil databases (Amini et al. 2005; Liao et al. 2014), and is used as an input in soil and environmental models (Keller et al. 2001). CEC refers to the quantity of negative charges in soil (Jaremko and Kalembasa 2014). The negative charge may be pH dependent (soil organic matter) or permanent (some clay minerals) (Liao et al. 2014; Zolfaghari et al. 2016). Although CEC can be measured directly, its measurement is difficult and expensive. Pedotransfer functions (PTFs) provide an alternative by estimating CEC from more readily available soil data (Liao et al. 2014; Emamgolizadeh et al. 2015; Zolfaghari et al. 2016).

In recent years, various PTFs have been developed to estimate CEC from basic physical and chemical soil properties (McBratney et al. 2002; Amini et al. 2005; Kianpoor et al. 2012; Bayat et al. 2014; Liao et al. 2014). In most of these models, CEC is assumed to be a linear function of soil organic carbon and clay content (McBratney et al. 2002; Sarmadian and Taghizadeh Mehrjardi (2008); Kianpoor et al. 2012). Multiple linear regression (MLR) analysis is generally used to find the relevant coefficients in the model equations. Often, however, models developed for one region may not give adequate estimates for a different region (Wagner et al. 2001; Amini et al. 2005; Emamgolizadeh et al. 2015).

A recent approach to model PTFs is the use of artificial neural networks (ANNs). Artificial neural networks have been successfully employed to predict some soil properties that their measurement is difficult (Minasny and McBratney 2002; Amini et al. 2005; Bayat et al. 2014; Emamgolizadeh et al. 2015). An advantage of using ANNs is that no specific type of function needs to be assumed a priori to model the relationship between inputs and outputs. The optimum relation that links input data to output data is obtained through a training procedure. ANN Models are generally expected to be superior to MLR models because of their greater feasibility (Amini et al. 2005; Bayat et al. 2014; Emamgolizadeh et al. 2015). A type of artificial neural network known as multi-layer perceptron (MLP), which uses a back-propagation training algorithm, is usually used for generating PTFs (Minasny and McBratney 2002; Amini et al. 2005; Sarmadian and Taghizadeh Mehrjardi (2008); Lake et al. 2009;Keshavarzia and Sarmadiana 2010; Yilmaz and Kaynar 2011; Kianpoor et al. 2012; Emamgolizadeh et al. 2015). This network uses neurons whose output is a function of a weighted sum of the inputs.

Several attempts have been conducted in relation to modeling various soil physiochemical parameters by means of different artificial intelligence-based model techniques such as those done for modeling of the daily and hourly behavior of runoff (Aqil et al. 2007), estimation of soil erosion and nutrient concentrations in runoff (Kim and Gilley 2008), modeling of Pb(II) adsorption from aqueous solution (Yetilmezsoy and Demirel 2008), to determine the clay dispersibility (Zorluer et al. 2010), estimating the grout ability of granular soils (Tekin and Akbas 2011), prediction of swell potential of clayey soils (Yilmaz and Kaynar 2011), prediction of soil water retention curve (Abbasi et al. 2011), land suitability evaluation (Keshavarzi et al. 2011), estimating wet soil aggregate stability (Besalatpour et al. 2013) and etc. Some studies also have been considered capability of soft computing techniques for prediction modeling soil CEC such as those conducted by Amini et al. (2005); Sarmadian and Taghizadeh Mehrjardi (2008); Tang et al. (2009); Sarmadian et al. (2013); Kianpoor et al. (2012), Keshavarzi et al. (2012), Liao et al. (2014), Bayat et al. (2014), Emamgolizadeh et al. (2015); Zolfaghari et al. (2016). The findings of these researchers demonstrated that PTFs developed through artificial intelligence-based modeling techniques were more efficient than the regression ones to predict the CEC. In spite of, few studies focused on developing PTFs by means of adaptive neuro-fuzzy inference system for prediction of CEC.

The objectives of this study were to develop suitable artificial neural network for estimation of CEC in Guilan region soils located in northern Iran and comparing artificial neural network with regression and adaptive neuro-fuzzy inference system models that have been developed for these soils.

Materials and methods

Study area and data collection

This research was carried out in paddy soils of Guilan province. The study area is located between 49°, 31′ to 49°, 45′E longitude and 37°, 7′ to 37°, 27′N latitude in north of Guilan Province, the southern coast of Caspian Sea, Northern Iran (Fig. 1). Region climate is very humid with annual precipitation mean 1293.6 mm and annual temperature mean 15.8 °C. The region soils moisture and temperature regimes are Aquic, Udic and Thermic, respectively, and soils parent materials are derived from river sediments. Soil series names of study area and their distribution are presented in Table 1 and Fig. 1, respectively. All soil profiles were deep expecting 10 and 11 soil series. Texture of soils was light to heavy in different soil series (Fig. 2).

The determination of chemical and physical properties was carried out on 171 soil samples collected from various horizons of 120 soil profiles. Using profile description and laboratory analysis of soil samples, all the studied soils

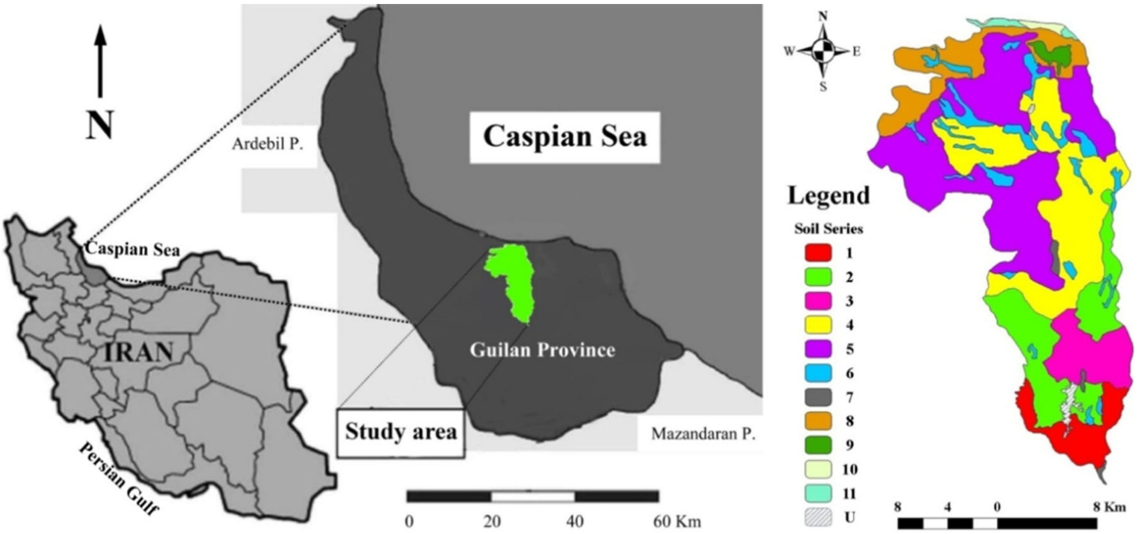


Fig. 1 Study area location and soil types map of area

Table 1 Soil series names of the study area with surface soil texture

Series no.	Characteristics	Top soil texture (0–30 cm)
1	Fine, mixed, active, thermic Anthraquic Eutrudepts	Very heavy: C
2	Fine, mixed, active, thermic Typic Endoaquepts	Heavy: SiCL
3	Fine, mixed, superactive, thermic Fluventic Endoaquepts	Very heavy: C
4	Fine, mixed, active, calcareous, thermic Typic Endoaquepts	Very heavy: C
5	Fine, mixed, active, thermic Fluventic Endoaquepts	Very heavy: C
6	Fine loamy, mixed, active, thermic Fluvaquentic Eutrudepts	Heavy: CL
7	Fine loamy, mixed, superactive, thermic Typic Udifluvents	Medium: SCL
8	Fine loamy, mixed, active, thermic Typic Endoaquepts	Medium: SC
9	Fine, mixed, active, thermic Mollic Epiaquepts	Very heavy: SiC
10	Mixed, thermic Typic Psammaquents	Light: LS
11	Fine loamy, mixed, superactive, thermic Typic Fluvaquents	Medium: L

Texture symbols: *C* clay, *SiCL* silty clay loam, *CL* clay loam, *SCL* sandy clay loam, *SC* sandy clay, *SiC* silty clay, *LS* loamy sand, *L* loam

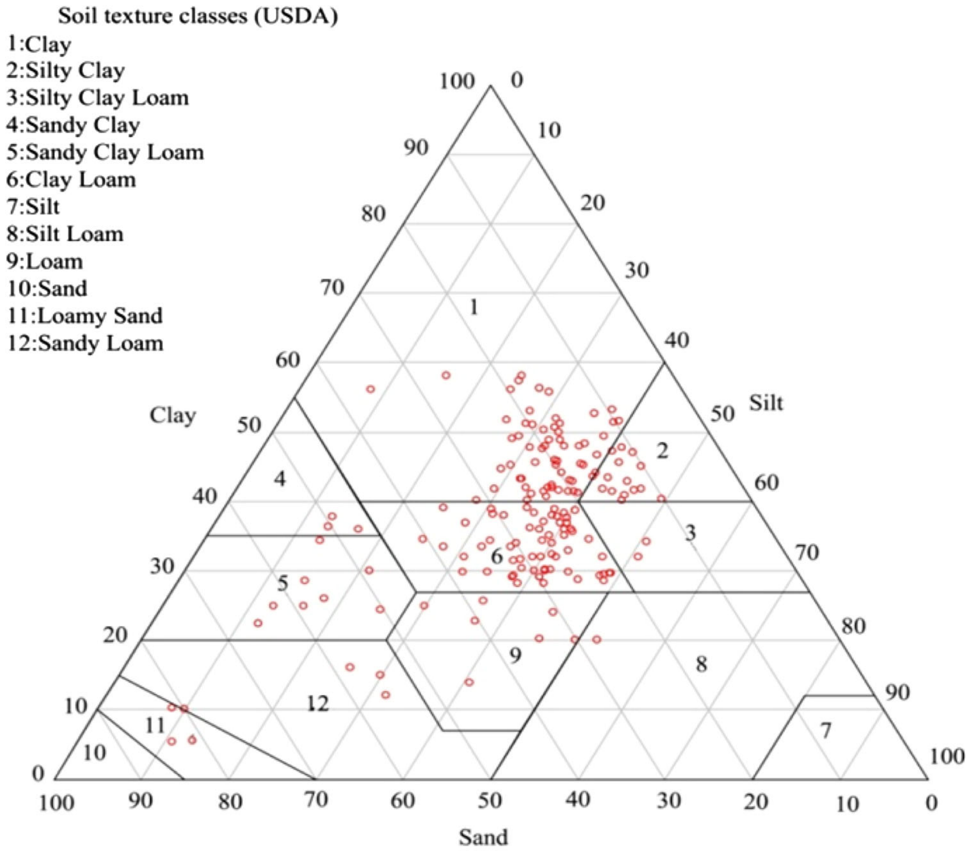
were classified as Entisols and Inceptisols on the basis of Soil Survey Staff (2014b). The soil properties measured for this research were organic carbon (OC), pH, calcium carbonate, CEC and soil texture fractions including clay, sand and silt content. The following analytical methods were employed to measure each of parameters for this study: organic carbon content was determined using Walkley–Black method (Nelson and Sommers 1982), Particle size distribution using pipette method (Soil Survey Staff 2014a), CEC using sodium acetate (pH 8.2), pH by pH meter in ratio 1:2 soil with water and $CaCO_3$ using titration method (Soil Survey Staff 2014a). The results of determinations were used as input variables to develop the CEC estimation models. The data sets were divided into two subsets. One subset was used for generating PTFs and calibrating PTFs from the literature, and the other subset was used for testing the models. The division was based on stratified random sampling by sorting the data based on CEC, stratifying the data into 10 CEC groups, and randomly selecting 25 % of the data from each group for testing. The remaining 75 % of the data were used for calibration and this division carried out as statistical characteristics of soil properties (e.g. min, max, and etc.) were similar in two subsets.

Prediction methods

Multiple linear and non-linear regression model

The general purpose of multiple regressions is to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable. Multiple regressions are the most common method used in development PTFs.

Fig. 2 Textural distribution of both training and testing data sets on the USDA soil texture triangle



Artificial neural networks (ANNs)

Artificial neural networks (ANNs) are a form of artificial intelligence, which, by means of their architecture, attempt is made to simulate the biological structure of the human brain and nervous system (Amini et al. 2005). In this study, developed ANN model was multi-layer perceptron which is the most commonly-used neural network structure in ecological modeling and soil science (Agyare et al. 2007; Besalatpour et al. 2013; Emamgolizadeh et al. 2015). The MLP algorithm developed in this research is a feed-forward back-propagation network (FFBP) model. There are two input elements, %Clay and %OC, and one output element, CEC, so that the MLP architecture is *2-m-1*, where m represents the number of hidden neurons. A schematic diagram of the network is given in Fig. 3. Assume P is a (d × n) rescaled input matrix where the rows consist of elements (i.e. clay and OC) and the columns are the samples. Initially, we calculate a linear combination, a_j, of the weighted input elements, P_i, plus a constant bias, $w_{jO}^{(h)}$, expressed as:

$$a_j = \sum_{i=1}^{d} w_{ji}^{(h)} P_i + w_{jo}^{(h)},\ j = 1,\ldots,m \text{ and } i = 1,\ldots,d \qquad (1)$$

where d is the number of elements, m is the number of neurons, and $w_{ji}^{(h)}$ denotes the weights given to the input i of the neuron j in the hidden layer. The matrix, a_j, is then activated by a tangent sigmoid function, f, to produce the output of the hidden layer, Z_j:

$$Z_j = f(a_j) = -1 + \left[2/\left(1 + exp(-2a_j)\right)\right] \qquad (2)$$

In the output layer (Fig. 3), the outputs of the hidden layer are summed linearly to produce CEC estimates:

$$CEC_{(predicted)} = \sum_{j=1}^{m} w_j^{(o)} Z_j + w_o^{(o)} \qquad (3)$$

The above procedure is repeated for every sample, i.e. n times. The weights in the above equations are adjustable parameters of the network and are optimized

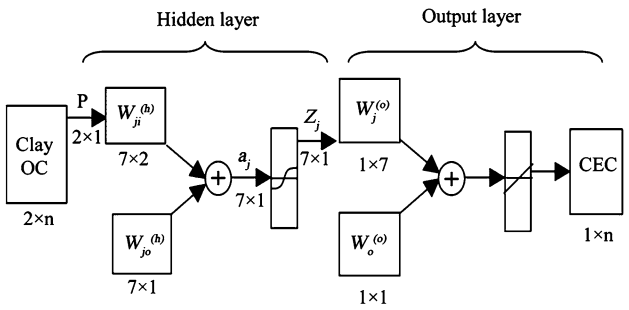


Fig. 3 A schematic structure of the feed forward back-propagation neural network

during the network training procedure. The commonly used objective function in training is the mean squared error (MSE) typically specified as:

$$MSE = \frac{1}{n}\sum_{k=1}^{n}\left(CEC_{predicted} - CEC_{measured}\right)^2, k = 1, \ldots, n. \tag{4}$$

Error minimization can be obtained by a number of procedures. Frequently, the Levenberg–Marquardt (More 1977) algorithm is used in feed-forward networks (Schaap et al. 1998). A problem that usually occurs during network training is over-fitting or overtraining, which means that the network learns to work well for the training inputs, but not well enough for a test data set. To avoid overtraining, Amini et al. (2005) proposed a regularized objective function, MSE_{Reg}, in which the sum of network weights is added to the MSE:

$$MSE_{Reg} = \gamma MSE + (1 - \gamma)MSW \tag{5}$$

where γ is a performance ratio calculated by means of the Bayesian regularization in combination with the Levenberg–Marquardt algorithm and MSW is the mean of the squared weights and biases (Amini et al. 2005). When the data set is small and you are training function approximation networks, Bayesian regularization provides better generalization performance than early stopping. This is because Bayesian regularization does not require that a validation data set be separate from the training data set; it uses all the data (Help of MATLAB R2015b software 2015). For this purpose, we used "create network or data toolbox" of MATLAB software which training, adaption learning, performance and transfer functions were Bayesian regularization, gradient descent, MSE_{Reg} and tangent sigmoid, respectively.

MATLAB R2015b software (2015) was used to develop PTFs for predicting CEC by means of ANN model. In order to this end, all data set were first normalized between 0.1 and 0.9 to achieve effective network training. Luk et al. (2000) stated that neural networks trained on normalized data, achieve better performance and faster convergence in general, although the advantages diminish as network and sample size become large. Normalizing the data set was done through in two stage: (1) Pre-processing: The input (clay and OC) and output (CEC) data for training and test data sets were initially rescaled to fall within the range of [0.1, 0.9] by the transfer function (Help of MATLAB R2015b software 2015):

$$P_{norm} = [0.8 \times ((P_i - P_{min})/(P_{max} - P_{min}))] + 0.1 \tag{6}$$

where P_{norm} is the rescaled input matrix, P_i is the input matrix, and P_{min} and P_{max} are two vectors containing the minimum and the maximum values of the input matrix, respectively. The output (CEC) of the network is also rescaled by using its minimum and maximum values. (2) Post-processing: To back-transform the results of the network we used the following equation:

$$P_i = [1.25 \times (P_{norm} - 0.1)/(P_{max} - P_{min})] + P_{min} \tag{7}$$

Adaptive neuro-fuzzy inference system (ANFIS) model

In ANFIS, fuzzy rule bases are combined with neural networks to train the system using experimental data and obtain appropriate membership functions for process prediction and control (Lertworasirikul 2008; Besalatpour et al. 2013). Takagi–Sugeno-Kang (TSK) model (Takagi and Sugeno 1985) that is one of the most frequently-used precise fuzzy models was used in the current study to predict soil CEC. In order to simplify, it is assumed that the inference system has two input variables x and y as each variable has two fuzzy subsets. A typical rule set with two fuzzy if–then rule set for a first-order Sugeno fuzzy model can be defined as Eqs. (8) and (9):

$$\text{Rule 1: If } x \text{ is } A_1 \text{ and } y \text{ is } B_1 \text{ Then } f_1 = p_1x + q_1y + r_1 \tag{8}$$

$$\text{Rule 2: If } x \text{ is } A_2 \text{ and } y \text{ is } B_2 \text{ Then } f_2 = p_2\text{x} + q_2y + r_2 \tag{9}$$

where A_1, A_2 and B_1, B_2 are the membership functions for inputs x and y respectively, p_1, q_1, r_1 and p_2, q_2, r_2 are the parameters of the output function. The corresponding equivalent ANFIS architecture for two input variable first-order Sugeno-fuzzy model with two rules is illustrated in Fig. 4a. The general architecture of ANFIS consists of five layers, namely, fuzzy, product, normalized, defuzzy and output layer is depicted in Fig. 4b. In this architecture, the circular nodes represent nodes that are fixed, whereas the square nodes are nodes that have parameters to be learnt (Yilmaz and Kaynar 2011).

Layer 1 Every node in this layer is represented by a square node including a node function. The node function employed by each node determines the membership relation between the input and output functions.

Layer 2 every node in this layer is a fixed (circle) labeled *II* node and its output is produced by signals obtained from layer 1.

Layer 3 every node in this layer is a fixed (circle) node labeled *N*. The nodes normalize the firing strength by calculating the ratio of firing strength for this node to the sum of all the firing strengths.

Layer 4 Every node in this layer is represented by a square node including a node function.

Layer 5 The single node in this layer is a fixed (circle) node labeled $\sum$ that computes the overall output as the summation of all incoming signals.

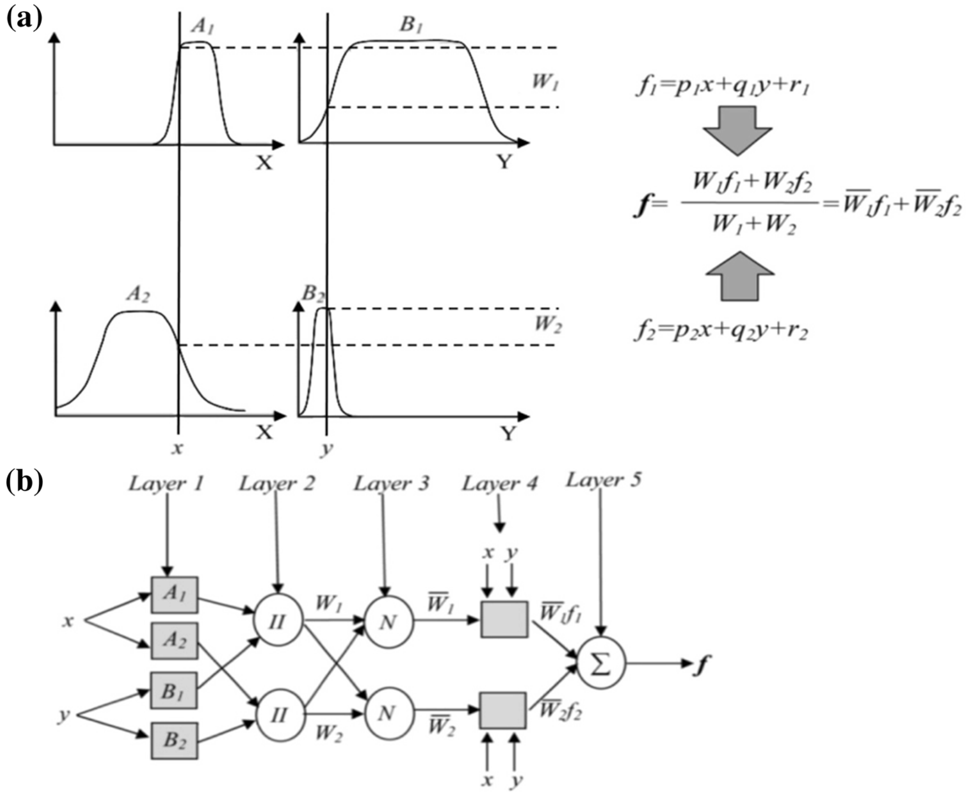


Fig. 4 **a** Two input first-order Sugeno-fuzzy model with two rules and **b** equivalent adaptive neuro-fuzzy inference system architecture

Performance evaluation criteria

Four different types of standard statistical performance evaluation criteria were used to control the accuracy of the prediction capacity of the models developed. These are root mean square error (RMSE), the determination coefficient (R^2), mean bias error (MBE) and relative improvement (RI). Performance evaluation criteria used in the current study can be calculated using following equations:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (10)$$

$$R^2 = 1 - \left[\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right)\Big/\left(\sum_{i=1}^{n}(y_i - \bar{y}_i)^2\right)\right] \quad (11)$$

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \quad (12)$$

$$RI = [(RMSE_{Reg} - RMSE_M)/RMSE_{Reg}] \times 100 \quad (13)$$

where y_i denotes the measured value, $\hat{y}_i$ is the predicted value, $\bar{y}_i$ is the average of the measured value, and n is the total number of observations. The MBE characterizes the mean difference between the calculated and measured data; hence, it is a criterion of systematic error in the model fitting. Negative and positive values of MBE indicate under and over estimation of PTFs for a given parameter respectively. $RMSE_{Reg}$ is root mean square error of regression model and $RMSE_M$ is root mean square error of other models (Bayat et al. 2014).

Results

Data summary statistics

Pertinent statistics of the soil properties used to calibrate and test the newly developed models are given in Table 2. The correlation coefficients between variables are given in Table 3. The correlations between CEC and soil OC (r = 0.63) and between CEC and clay content (r = 0.82) had the most value and were positive significant in 0.01 level in comparison with the other properties. Therefore, clay and organic carbon content were used for prediction of CEC. The coefficient of variation (CV) of the soil organic carbon content showed more variability than those of the soil clay percentage and CEC, being about three times as large as the other properties (Table 2). This large variation in OC is due largely to the variability in manure and compost applications as fertilizer and return of rice plant residuals and soil amendments in the study area.

Multiple linear and non-linear regressions (MLR and MNLR)

Developing PTFs using MLR and MNLR models for predicting soil CEC in study area were done by means of SPSS 24 software (2016) and above-mentioned physiochemical soil properties were used as independent variables. In the regression analysis, normalizing the data distribution is one of the primary assumptions that have to be carried out. Therefore, the normality of the data was evaluated using the Kolmogorov–Smirnov method. All data had a normal distribution. After normalizing test data, multiple linear regression function was derived for training data set through stepwise method. In this method, all data were first inserted as input data and subsequently, the data that were significantly less effective on output parameter were eliminated. MLR model was derived among CEC, OC and clay content properties (Eq. 14). It was found that the developed equations through MLR model among CEC and input variables were not statistically strong enough to establish significant models by traditional statistical models, because few numbers of inputs had high correlation with CEC. However, since the accuracy of pedotransfer function models depends on the number of inputs, while increasing the number of inputs will decrease the accuracy of the estimations (Amini et al. 2005). OC and clay were used for developing non-linear regression model. Different types of models include power, exponential, cubic and etc. were developed for non-linear regression. Finally, the best linear and non-linear regression equations that were derived for training data set were as Eqs. (14) and (15) that variance analysis result of multiple linear and non-linear regression models was mentioned in Table 4:

$$CEC = 4.263 + 0.455\,Clay + 1.097\,OC \quad R^2 = 0.77 \tag{14}$$

$$CEC = 0.55 + 0.64\,Clay^{0.97} + 0.55\,OC^{1.26} \quad R^2 = 0.79 \tag{15}$$

From the numerous available PTFs derived to predict CEC we selected only those regression models that used OC and clay as independent variables and had a coefficient of determination, R^2, greater than 0.5. The selected PTFs were calibrated for the study region using a generalized least squares procedure with a subset of training data (Table 2). The models and their evaluation criteria amounts are given in Table 5. The R^2 and RMSE values of models showed that regression models of current study were the most accurate one.

Optimization of artificial neural network model

We used feed-forward back-propagation neural network in this study. We constructed one network that used organic carbon and clay content as inputs. Because, former researchers such as Amini et al. (2005), Sarmadian and Taghizadeh Mehrjardi (2008). Sarmadian et al. (2013) found that these inputs had the best results. Also, these inputs had the most correlation coefficient with CEC in current study (Table 3). Finding the optimum number of hidden neurons in the hidden layer is an important step in developing FFBP networks. In neural network design, too many hidden units cause over-fitting, while too few hidden

Table 2 Statistics of the training and testing data sets

Parameter	CEC	OC	Clay	Silt	Sand	$CaCO_3$	pH
Training data (n = 131)							
Mean	23.63	1.79	38.22	37.03	26.48	7.50	7.31
SD	4.95	1.30	9.04	8.05	12.13	1.35	0.45
CV	20.9	72.6	23.6	21.7	45.8	18	6.15
Min	8.56	0.04	5.4	8.30	8.80	5.12	6.14
Max	36.20	7.96	58	52.30	83.70	10.12	8.00
Skewness	0.51	0.49	−0.43	0.31	0.47	0.07	0.43
*Asymp.Sig.	0.36	0.57	0.78	0.85	0.65	0.78	0.69
Testing data (n = 40)							
Mean	23.26	1.98	38.16	38.21	26.25	7.41	7.34
SD	3.97	1.32	6.81	7.98	11.35	1.45	0.48
CV	17.1	66.7	17.8	20.8	43.2	19.6	6.53
Min	16.10	0.2	21.20	8.32	8.61	5.56	6.21
Max	30.20	5.17	52.60	51.8	82.5	9.58	7.89
Skewness	0.01	0.36	0.09	−0.38	0.41	0.04	0.28
Asymp.Sig.	0.89	0.33	0.94	0.78	0.52	0.78	0.58

* Asymp.Sig.: Kolmogorov–Smirnov test index for normal distribution, that should be greater than 0.05. CEC measured in $cmol(+)kg^{-1}$ and OC, clay, silt, sand, $CaCO_3$ in percentage (%)

Table 3 Correlation coefficients of the measured soil properties

	CEC	OC	Clay	Silt	Sand	$CaCO_3$	pH
CEC	1						
OC	0.63**	1					
Clay	0.82**	0.25*	1				
Silt	0.21*	0.19*	0.16	1			
Sand	−0.53**	−0.19*	−0.76**	−0.76**	1		
$CaCO_3$	−0.15	0.03	−0.03	−0.14	0.12	1	
pH	0.21*	0.23*	0.31*	0.09	−0.09	−0.19*	1

* Correlation is significant at the 0.05 level

** Correlation is significant at the 0.01 level

Table 4 Variance analysis result of multiple linear and non-linear regression models

Model	Source	Sum of squares	df	Mean square	F	Sig.
Linear	Regression	2477.12	2	1238.56	220.77	.000
	Residual	718.46	128	5.61		
	Total	3195.58	130			
Non-linear	Regression	75504.63	1	75504.63	14632.68	.000
	Residual	671.07	130	5.16		
	Uncorrected total	76175.70	131			
	Corrected total	3195.58	130			

Table 5 Selected pedotransfer functions and their calibration coefficients

Model	References	PTF model	R^2	RMSE
M1	McBratney et al. (2002)	CEC = 6.9 + 0.1 clay + 0.16 (clay × OC)	0.55	2.295
M2	Sarmadian and Taghizadeh Mehrjardi (2008)	CEC = 1.91 + 0.318 clay + 3.96 OC	0.52	2.521
M3	Lake et al. (2009)	CEC = 12.6 + 2.03 clay + 0.1 OC	0.58	1.852
M4	Keshavarzia and Sarmadiana (2010)	CEC = 10.6 + 0.19 clay + 1.37 OC	0.62	1.729
M5	Current study (linear model)	CEC = 4.263 + 0.455 clay + 1.097 OC	0.77	1.293
M6	Current study (non-linear model)	CEC = 0.55 + 0.64 $\text{clay}^{0.97}$ + 0.55 $\text{OC}^{1.26}$	0.79	1.093

units cause under fitting. To find the optimum number of hidden units, the RMSEs of the network with two inputs (OC and clay) were plotted versus the number of hidden units (Fig. 5), and number of hidden units equal 7 had lower RMSE therefore it be selected.

The objective function without regularization (not shown) produced numerous local minima and fluctuated greatly as the number of hidden units increased. The regularized objective function, Eq. (6), showed a more stable response with respect to the number of hidden units as the RMSE of the training and testing procedures decreased gradually as the number of hidden units increased to 7 (Fig. 5). There are a number of advantages in using the Bayesian regularization algorithm (BRA) over others. One advantage is that it is model-driven rather than data-driven; owing to its Bayesian principles as opposed to maximum likelihood principles. Another advantage of BRA is that of pruning. The weight penalty term that is added to the algorithm means that, as long as sufficient hidden neurons are supplied, the BRA will automatically prune the ANN to the optimum architecture and over-fitting is avoided (Amini et al. 2005). For this reason, as illustrated by Fig. 5, adding more hidden neurons do not improve the model. After repeated experiments, a persistent minimum value of RMSE occurred at the hidden unit value of 7 with two inputs (Fig. 5), suggesting that pruning occurred above 7 hidden units. Therefore, we used a FFBP network containing 7 hidden units ($FFBP_{7H}$), with tangent sigmoid transfer function, Bayesian regularization training function and gradient descent adaptation learning function for further analysis. The weights for the $FFBP_{7H}$ model are given in Table 6. Comparison of results obtained from current study with Sarmadian and Taghizadeh Mehrjardi (2008); Tang et al. (2009); Lake et al. (2009); Keshavarzia

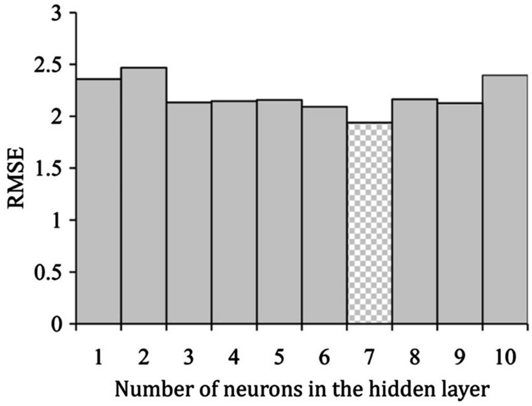


Fig. 5 RMSE versus number of neuron in hidden layer in FFBP network for selective suitable number of neurons

Table 6 The weights used for the two FFBP networks with 7 neurons

	Hidden neurons (j = 1, …, 7)						
	1	2	3	4	5	6	7
$w_{j1}^{(h)}$	3.076	3.720	1.653	−3.070	1.877	3.707	3.506
$w_{j2}^{(h)}$	1.736	0.742	−3.429	−1.318	3.241	0.979	−1.351
$w_{j0}^{(h)}$	−3.863	−2.805	1.438	1.132	1.571	2.136	−3.505
$w_{j}^{(o)}$	0.416	−1.260	0.508	−0.767	0.245	0.209	0.580
$w_{0}^{(o)}$	−0.367						

$w_{j1}^{(h)}$, $w_{j2}^{(h)}$ and $w_{j0}^{(h)}$ are the weights in the hidden layer for clay, organic carbon content and bias respectively, $w_{j}^{(o)}$ and $w_{0}^{(0)}$ are the weights in the output layer

and Sarmadiana (2010); Kianpoor et al. (2012); Sarmadian et al. (2013); Bayat et al. (2014); Emamgolizadeh et al. (2015) researches showed that using Bayesian regularization algorithm for model learning in this study caused to increase accuracy of artificial neural network for estimation of CEC. Also, Amini et al. (2005), increased accuracy of ANN model using Bayesian regularization learning algorithm that our study result was in agreement with their research results, too.

Adaptive neuro-fuzzy inference system (ANFIS)

In this study, ANFIS model was also applied for predicting CEC using the same normalized data that were used for ANN model. In the ANFIS system, each input parameter might be clustered into several class values in layer 1 to build up fuzzy rules and each fuzzy rule would be constructed using two or more membership functions in layer 2. Several methods have been proposed to classify the input data and to make the rules, among which the most widespread are grid partition and subtractive fuzzy clustering (Aqil et al. 2007; Ertunc and Hosoz 2008; Yetilmezsoy et al. 2011; Kianpoor et al. 2012). In this study, grid partition was taken in consideration. Then Psigmoid membership function and their numbers for input parameters and linear membership function for output parameter were selected and so, fuzzy inference system (FIS) was generated. For training FIS, hybrid algorithm was applied. In this way, epoch 40 had the most optimal result with minimum error. After the FIS was trained, validation of the model using a testing data was carried out. Different parameter types and their values used for training ANFIS can be seen in Table 7. The descriptive performance of the ANFIS model for the test dataset and the related statistical evolutionary results are given in Table 8. The values of 0.82, 1.184, 0.218 and 25.7 for R^2, RMSE, MBE and RI parameters, respectively, for ANFIS testing stage, while regression and ANFIS efficiency were less than feed-forward back-propagation network model. Comparison of trained ANFIS model in this study with trained ANFIS in Kianpoor et al. (2012) and Keshavarzi et al. (2012) researches showed that accuracy of ANFIS was high in current study, because, we used grid partition for classification of input data and making the rules, However, they used subtractive fuzzy clustering. Therefore, using grid partition caused to increase accuracy of training in our research. Yilmaz and Kaynar (2011) and Vafakhah et al. (2014) used grid partition and hybrid algorithm for FIS generation and training, respectively, in ANFIS model and reported high accuracy for model training. And so, our result was in agreement with them.

Table 7 Different parameter types and their values used for training ANFIS

ANFIS parameter type	Value
MF type	Psigmoid
Number of MFs	5
Output function	Linear
Number of linear parameters	75
Number of nonlinear parameters	40
Total number of parameters	115
Number of training data pairs	131
Number of testing data pairs	40
Learning algorithm	Hybrid
Epoch	40

Discussion

After determining regression equations, in order to evaluate the accuracy of MLR and MNLR models, the results of these models were compared with experimental data. In

Table 8 Test results of the regression, neural network and adaptive neuro-fuzzy inference system

Model	R^2	RMSE	MBE	RI
Linear regression	0.68	1.593	−0.328	0
Non-linear regression	0.73	1.364	0.286	14.3
ANFIS	0.82	1.184	0.218	25.7
ANN (FFBP)	0.96	0.314	0.106	80.3

fact, the coefficient of determination (R^2) between the measured and predicted values is a good indicator to check the prediction performance of the model (Gokceoglu and Zorlu 2004; Kianpoor et al. 2012). The obtained values of R^2, RMSE, MBE and RI using MLR and MNLR are shown in Table 8. For test dataset, the R^2, RMSE and MBE values have been obtained 0.68, 1.593 and −0.328 for linear regression and 0.73, 1.364 and 0.286 for non-linear regression, respectively. Results showed that MNLR model have high accuracy with regard to MLR and this shows relationship between CEC and soil properties such as organic carbon and clay is non-linear and complex. MLR model result is in contrast with the results of Yilmaz et al. (2012) and Kianpoor et al. (2012). However, obtained results had agreement with those reported by McBratney et al. (2002); Amini et al. (2005); Sarmadian and Taghizadeh Mehrjardi (2008); Bayat et al. (2014); Emamgolizadeh et al. (2015). Their results showed high correlation coefficient for predicting the soil CEC by means of multiple linear regression models. As above mentioned; the more inputs will result in the less accuracy of the estimation (Amini et al. 2005) and this point explains their results. Input data in McBratney et al. (2002); Amini et al. (2005) and Sarmadian and Taghizadeh Mehrjardi (2008) studies were clay and organic carbon.

The test data set was used to evaluate the performance of the MLR, MNLR, neural network model and ANFIS for predicting CEC. The statistical results of the comparisons are given in Table 8, which shows that the neural network model had larger R^2 value than the regression and ANFIS models. This is in line with the work done by Yilmaz and Kaynar (2011); Kianpoor et al. (2012); Bayat et al. (2014). Their findings demonstrated that prediction performances of the FFBP model had higher accuracy than both multiple regression equations and adaptive neuro-fuzzy inference system for predicting swell potential of clayey soil and CEC, respectively. The MBE values indicated that the artificial neural network and ANFIS models had overestimated the CEC. This overestimation was however small, especially for the FFBP model. The smallest RMSE was produced by the $FFBP_{7H}$ model, while the largest RMSE was produced by the linear regression model, these results were in agreement with Amini et al. (2005); Kianpoor et al. (2012); Sarmadian et al. (2013). The relative improvement of the models was calculated using the linear regression model as a reference. The results in Table 8 show that the FFBP model had in general the largest relative improvement (RI) that was in agreement with Amini et al. (2005); Bayat et al. (2014). The scatter plots of the measured versus predicted CEC for the test data set are given in Fig. 6 for the prediction models, which we identified $FFBP_{7H}$ as being the best model for predicting CEC.

On the other hand, the proposed ANN model was, in general, more feasible than the ANFIS model in predicting CEC when the evaluation criteria are compared. The existing patterns and trends among the input variables and the output (CEC) are relatively complex and intricate. It appears that, the ANN model was more capable in extracting the existing patterns among the input variables and the output. Neural networks can extract the patterns and detect the trends that are too complex to be noticed by either humans or other computer techniques because of their remarkable ability to derive a general solution from complicated or imprecise data (Yilmaz and Kaynar 2011; Besalatpour et al. 2013; Bayat et al. (2014)). These artificial networks have the capability of learning from examples and are capable to solve intricate, nonlinear problems and problems which are very tedious to solve by conventional methods. In addition, when a data stream is analyzed using a neural network, it is possible to detect the important predictive patterns that are not previously apparent to a non-expert (Yilmaz and Kaynar 2011; Besalatpour et al. 2013). Finally, all these indicate that ANFIS approach may not always be a better choice for predicting soil CEC.

Conclusion

In this study, multiple linear and non-linear regression, artificial neural network models (feed-forward back-propagation network, FFBP) and adaptive neuro-fuzzy inference system were employed to develop a pedotransfer function for predicting soil CEC using available soil properties. The performance of the regression, neural network and ANFIS models was evaluated using a test data set. The newly developed FFBP neural network PTF with 7 hidden neurons predicted CEC better than the regression and ANFIS models and significantly improved the accuracy of the prediction by up to 80.3 %. The neural network models are in general more suitable for capturing the non-linearity of the relationship between variables. In this study, however, the relationship between CEC and clay and organic carbon appeared to be dominantly linear. Consequently, with the use of proposed ANNs especially, FFBP network, the performance of CEC condensers can be

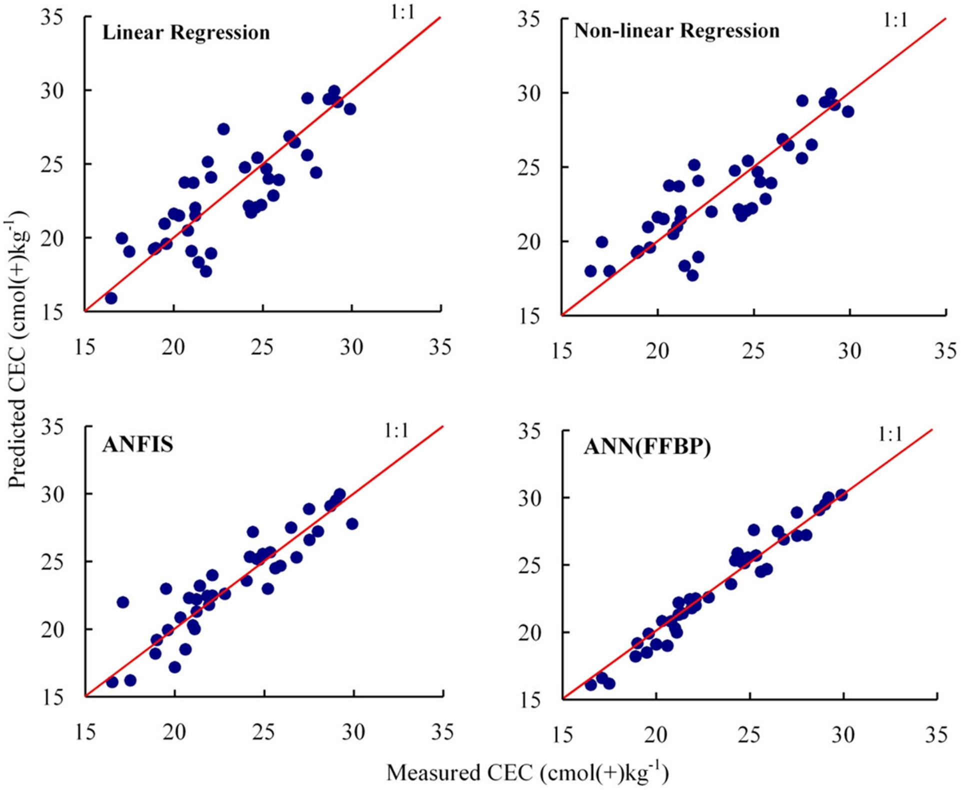


Fig. 6 The scatter plots of the measured versus predicted CEC for testing data using regression, ANFIS and FFBP$_{7H}$ network with two inputs (OC and clay)

determined by performing only a limited number of test operations, thus saving engineering effort, time and funds. Finally, using Bayesian regularization algorithm for model learning in FFBP and grid partition for classification of input data and making the rules in ANFIS model caused to increase the accuracy of these models, dramatically, for CEC prediction. We suggest that researchers use genetic algorithm for optimization of models and rules in future work.

Acknowledgments This research was supported by Guilan University, Iran. The authors thank for laboratory assistance of soil science department technicians in university of Guilan. The authors are grateful to anonymous reviewers who considerably improved the quality of the manuscript.

References

Abbasi Y, Ghanbarian Alavijeh B, Liaghat AM, Shorafa M (2011) Evaluation of pedotransfer functions for estimating soil water retention curve of saline and saline-alkali soils of Iran. Pedosphere 21(2):230–237. doi:10.1016/S1002-0160(11)60122-7

Agyare WA, Park SJ, Vlek PLG (2007) Artificial neural network estimation of saturated hydraulic conductivity. Vado Zone J 6:423–431. doi:10.2136/vzj2006.0131

Amini M, Abbaspour KC, Khademi H, Fathianpour N, Afyuni M, Schulin R (2005) Neural network models to predict cation exchange capacity in arid regions of Iran. Euro J Soil Sci 56:551–559. doi:10.1111/j.1365-2389.2005.0698.x

Aqil M, Kita I, Yano A, Nishiyama S (2007) A comparative study of artificial neural networks and neuro-fuzzy in continuous modeling of the daily and hourly behaviour of runoff. J Hydrol 337:22–34. doi:10.1016/j.jhydrol.2007.01.013

Bayat H, Davatgar N, Jalali M (2014) Prediction of CEC using fractal parameters by artificial neural networks. Int Agrophys 28:143–152. doi:10.2478/intag-2014-0002

Besalatpour AA, Ayoubi S, Hajabbasi MA, Mosaddeghi MR, Schulin R (2013) Estimating wet soil aggregate stability from easily available properties in a highly mountainous watershed. Catena 111:72–79. doi:10.1016/j.catena.2013.07.001

Emamgolizadeh 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 529(3):1590–1600. doi:10.1016/j.jhydrol.2015.08.025

Ertunc HM, Hosoz M (2008) Comparative analysis of an evaporative condenser using artificial neural network and adaptive neuro-fuzzy inference system. Int J Refrig 31(8):1426–1436. doi:10.1016/j.ijrefrig.2008.03.007

Gokceoglu C, Zorlu K (2004) A fuzzy model to predict the uniaxial compressive strength and the modulus of elasticity of a problematic rock. Eng Appl Artif Int 17:61–72. doi:10.1016/j.engappai.2003.11.006

Jaremko D, Kalembasa D (2014) A comparison of methods for the determination of cation exchange capacity of soils. Ecol Chem Eng S 21(3):487–498

Keller A, Von Steiger B, vander Zee ST, Schulin R (2001). A stochastic empirical model for regional heavy metal balances in agroecosystems. J Environ Qual 30(6):1976–1989. http://www.ncbi.nlm.nih.gov/pubmed/11790004

Keshavarzi A, Sarmadian F, Ahmadi A (2011) Spatially-based model of land suitability analysis using Block Kriging. Aust J Crop Sci

5(12):1533–1541. http://search.informit.com.au/documentSummary;dn=005523601697437;res=IELHSS

Keshavarzi A, Sarmadian F, Rahmani A, Ahmadi A, Labbafi R, Iqbal MA (2012). Fuzzy clustering analysis for modeling of soil cation exchange capacity. Aust J Agric Eng 3(1):27–33. http://search.informit.com.au/documentSummary;dn=339426446644381;res=IELENG

Keshavarzia A, Sarmadiana F (2010). Comparison of artificial neural network and multivariate regression methods in prediction of soil cation exchange capacity (CDepartment of Soil Sciencease study: Ziaran region). DESERT 15:167–174. https://journals.ut.ac.ir/article_23014.html

Kianpoor KY, Rezaie AR, Amerikhah H, Sami M (2012). Comparison of multiple linear regressions and artificial intelligence-based modeling techniques for prediction the soil cation exchange capacity of Aridisols and Entisols in a semiarid region. Aust J Agric Eng 3(2):39–46. http://search.informit.com.au/documentSummary;dn=722203574735923;res=IELENG

Kim M, Gilley JE (2008) Artificial Neural Network estimation of soil erosion and nutrient concentrations in runoff from land application areas. Comput Electron Agric 64:268–275. doi:10.1016/j.compag.2008.05.021

Lake HR, Akbarzadeh A, Taghizadeh Mehrjardi R 2009. Development of pedotransfer functions (PTFs) to predict soil physico-chemical and hydrological characteristics in southern coastal zones of the Caspian Sea. J Ecol Nat Environ 1(7):160–172. http://www.academicjournals.org/JENE

Lertworasirikul S (2008) Drying kinetics of semi-finished cassava crackers: a comparative study. LWT-Food Sci Technol 41(8):1360–1371. doi:10.1016/j.lwt.2007.09.009

Liao K, Xu Sh, Wu J, Zhu Q, An L (2014) Using support vector machines to predict cation exchange capacity of different soil horizons in Qingdao City, China. J Plant Nutri Soil Sci 177(5):775–782. doi:10.1002/jpln.201300176

Luk KC, Ball JE, Sharma A (2000) A study of optimal model lag and spatial inputs to artificial neural network for rainfall forecasting. J Hydrol 227:56–65. doi:10.1016/S0022-1694(99)00165-1

MATLAB R2015b (2015) Software for technical computing and model-based design. The Math Works Inc, USA

McBratney AB, Minasny B, Cattle SR, Vervoort RW (2002) From pedotransfer functions to soil inference systems. Geoderma 109(1–2):41–73. doi:10.1016/S0016-7061(02)00139-8

Minasny B, McBratney AB (2002) The neuro-m methods for fitting neural network parametric pedotransfer functions. Soil Sci Soc Am J 66:352–361. doi:10.2136/sssaj2002.3520

More J (1977) The Levenberg–Marquardt algorithm: implementation and theory. In: Watson GA (ed) Numerical analysis. Springer-Verlag, Heidelberg, pp 105–116. doi:10.1007/BFb0067700

Nelson DW, Sommer LE (1982). Total carbon, organic carbon and organic matter. In: Page AL (ed), Methods of soil analysis. Part 2. Chemical and microbiological properties. Monograph no, 9. ASA, Madison, pp 539–579

Sarmadian F, Taghizadeh Mehrjardi R (2008). Modeling of some soil properties using artificial neural network and multivariate regression in Gorgan province, north of Iran. Glob J Environ Res 2(1): 30–35. http://www.idosi.org/

Sarmadian F, Azimi S, Keshavarzi A, Ahmadi A (2013) Neural computing model for prediction of soil cation exchange capacity: a data mining approach. Inter J Agron Plant Prod 4(7):1706–1712. https://www.cabdirect.org/cabdirect/abstract/20133241195

Schaap MG, Leij FJ, van Genuchten MTh (1998) Neural network analysis for hierarchical prediction of soil hydraulic properties. Soil Sci Soc Am J 62:847–855. doi:10.2136/sssaj1998.03615995006200040001x

Soil Survey Staff (2014a). Kellogg soil survey laboratory methods manual. Soil survey investigations report no. 42, Version 5.In: Burt R, Soil Survey Staff (ed) United States Department of Agriculture, Natural Resources Conservation Service

Soil Survey Staff (2014b) Keys to soil taxonomy, 12th edn. United States Department of Agriculture, Washington D.C

SPSS 24 (2016) Statistical analysis software (Standard Version). SPSS Inc., USA

Takagi T, Sugeno M (1985) Fuzzy identification of systems and its applications to modeling and control. IEEE Trans Syst Man Cybern 15:116–132. doi:10.1109/TSMC.1985.6313399

Tang L, Zeng G, Nourbakhsh F, Guoli L, Shen GL (2009). Artificial neural network approach for predicting cation exchange capacity in soil based on physico-chemical properties. Environ Eng Sci 26(1):137–146. http://online.liebertpub.com/doi/abs/10.1089/ees.2007.0238

Tekin E, Akbas SO (2011) Artificial neural networks approach for estimating the groutability of granular soils with cement-based grouts. Bull Eng Geol Environ 70:153–161. doi:10.1007/s10064-010-0295-x

Vafakhah M, Janizadeh S, Khosrobeigi Bozchaloei S (2014) Application of several data-driven techniques for rainfall-runoff modelling. ECOPERSIA. 2(1):455–469

Wagner B, Tarnawski VR, Hennings V, Muller U, Wessolek G, Plagge R (2001) Evaluation of pedotransfer functions for unsaturated soil hydraulic conductivity using an independent dataset. Geoderma 102:275–279. doi:10.1016/S0016-7061(01)00037-4

Yetilmezsoy K, Demirel S (2008) Artificial neural network (ANN) approach for modeling of Pb(II) adsorption from aqueous solution by Antep pistachio (Pistacia Vera L.) shells. Hazard Mater J 153:1288–1300. doi:10.1016/j.jhazmat.2007.09.092

Yetilmezsoy K, Fingas M, Fieldhouse B (2011) An adaptive neuro-fuzzy approach for modeling of water-in-oil emulsion formation. Colloids Surf A Physicochem Eng Aspect 389:50–62. doi:10.1016/j.colsurfa.2011.08.051

Yilmaz I, Kaynar O (2011) Multiple regression, ANN (RBF, MLP) and ANFIS models for prediction of swell potential of clayey soils. Expert Syst Appl 38:5958–5966. doi:10.1016/j.eswa.2010.11.027

Yilmaz I, Marschalko M, Bednarik M, Kaynar O, Fojtova L (2012) Neural computing models for prediction of permeability coefficient of coarse-grained soils. Neural Comput Applic 21(5):957–968. doi:10.1007/s00521-011-0535-4

Zolfaghari AA, Taghizadeh-Mehrjardi R, Moshki AR, Malone BP, Weldeyohannes AO, Sarmadian F, Yazdani MR (2016) Using the nonparametric k-nearest neighbor approach for predicting cation exchange capacity. Geoderma 265:111–119. doi:10.1016/j.geoderma.2015.11.012

Zorluer I, Icaga Y, Yurtcu S, Tosun H (2010) Application of a fuzzy rule-based method for the determination of clay dispersibility. Geoderma 160:189–196. doi:10.1016/j.geoderma.2010.09.017