Computers and Electronics in Agriculture 135 (2017) 242–251

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Modeling soil cation exchange capacity using soil parameters: Assessing the heuristic models

Jalal Shiri [a,*], Ali Keshavarzi [b], Ozgur Kisi [c], Ursula Iturraran-Viveros [d], Ali Bagherzadeh [e], Rouhollah Mousavi [b], Sepideh Karimi [a]

[a] Water Engineering Department, Faculty of Agriculture, University of Tabriz, Tabriz, Iran
[b] Laboratory of Remote Sensing and GIS, Department of Soil Science, University of Tehran, P.O. Box: 4111, Karaj 31587-77871, Iran
[c] Center for Interdisciplinary Research, International Black Sea University, Tbilisi, Georgia
[d] Departamento de Matematicas, Facultad de Ciencias, Universidad Nacional Autonoma de Mexico, Circuito Escolar, Cd. Universitaria, Coyoacán 04510, Ciudad de México, Mexico
[e] Department of Agriculture, Islamic Azad University, Mashhad Branch, Emamyeh Boulevard, P.O. Box: 91735-413, Mashhad, Iran

ARTICLE INFO

Article history:
Received 22 November 2016
Received in revised form 10 January 2017
Accepted 16 February 2017

Keywords:
Cation exchange capacity
Heuristic models
k-fold testing

ABSTRACT

Accurate knowledge about soil cation exchange capacity (CEC) is very important in land drainage and reclamation, groundwater pollution studies and modeling chemical characteristics of the agricultural lands. The present study aims at developing heuristic models, e.g. gene expression programming (GEP), neuro-fuzzy (NF), neural network (NN), and support vector machine (SVM) for modeling soil CEC using soil parameters. Soil characteristic data including soil physical parameters (e.g. silt, clay and sand content), organic carbon, and pH from two different sites in Iran were utilized to feed the applied heuristic models. The models were assessed through a k-fold test data set scanning procedures, so a complete scan of the possible train and test patterns was carried out at each site. Comparison of the models showed that the NF outperforms the other applied models in both studied sites. The obtained results revealed that the performance of the applied models fluctuated throughout the test stages and between two sites, so a reliable assessment of the model should consider a complete scan of the utilized data set, which will be a good option for preventing partially valid conclusions obtained from assessing the models based on a simple data set assignment.

1. Introduction

The soil cation exchange capacity (CEC) is the total exchangeable cations which may be hold in soil by electrostatic forces at a specific pH level (Bauer and Velde, 2014). The knowledge about CEC values is important in land drainage and reclamation as well as groundwater pollution studies (van Hoorn and van Alphen, 1994). It is one of the most important chemical characteristics of agricultural lands (Ghaemi et al., 2013). CEC influences the stability of soil structure, nutrient availability, soil pH and the soil's reaction to fertilizers and other ameliorants, as well as it and provides a buffer against soil acidification (Hazelton and Murphy, 2007). It is also used as a measure of soil fertility, nutrient retention capacity, and the capacity to protect groundwater from cation contamination (Robertson et al., 1999). Usually, heavy clay soils present higher magnitudes of CEC, expressing the higher availability of nutrients in these soils.

CEC is usually measured on the fine earth fraction (soil particles lower than 2 mm in size). In gravelly soils, the effective soil CEC as a whole is diluted, and if only the clay fraction is analyzed, the obtained CEC values will be higher than the actual field values. Measuring CEC includes washing the soil for removing excess salts and using an 'index ion' for determining the total positive charge in relation to original soil mass. This includes bringing the soil to a predetermined pH level before analysis. Further details about CEC measurement techniques might be found in e.g. Rengasamy and Churchman (1999) and Rayment and Higginson (1992). However, these methods are time consuming, laborious and expensive, especially in remote areas, e.g. Aridisols in Iran. Alternatively, heuristic data driven models [e.g. gene expression programming (GEP), neuro-fuzzy (NF) technique, neural networks (NN) and support vector machine (SVM)] which can relate the CEC to its influential parameters might be applied for simulating CEC. Were et al. (2015) compared different heuristic models for predicting soil organic carbon stocks across an Afromontane landscape and found the SVM as the superior model in this issue. Keshavarzi et al. (2015) applied neural-network for defining pedotransfer functions

* Corresponding author.
E-mail addresses: j_shiri2005@yahoo.com, jalalshiri@tabrizu.ac.ir (J. Shiri).

http://dx.doi.org/10.1016/j.compag.2017.02.016
0168-1699/

in estimating soil phosphorous. Keshavarzi et al. (2017) developed ANFIS-based subtractive clustering algorithm in estimating soil CEC through using soil and remotely sensed data in a semi-arid region of Iran. Emamgolizadeh et al. (2016) compared different heuristic models for predicting CEC and found that the multivariate adaptive regression splines (MARS) and GEP models are superior in this issue. Zolfaghari et al. (2016) applied k-nearest neighbor technique for predicting soil cation exchange capacity. All the reported literature have used single data set assignment for developing and testing the applied models, where the models are trained by using a portion of the whole data and tested using the remain data patterns. The present study, however, aims at assessing heuristic data driven approaches, namely GEP, NF, NN and SVM, in modeling soil CEC through k-fold testing. The necessary input variables of the models were identified by utilizing Gamma-test. This is the first attempt that compares the GEP, NF, NN and SVM methods accuracy in modeling soil CEC.

2. Materials and methods

2.1. Gene expression programming (GEP)

In contrast to common applications of classical regression models to estimate CEC indirectly based on other data (Bishop and McBratney, 2001; Park and Vlek, 2002; Triantafilis et al., 2011), GP (genetic programming) has not been exploited for this purpose, although it has shown much potential in similar applications (Johari et al., 2006; Makkeasorn et al., 2006; Parasuraman et al., 2007; Padarian et al., 2012).

GP-based models (Koza, 1992), utilize a "parse tree" structure for the search of their solutions. GP has the ability for creating an explicit formulations set that govern the studied phenomenon, to map the relationship(s) between the input-target parameters using different operators. Gene expression programming (GEP) is similar to GP, in a way that selects the best governing formulations based on fitness values and introduces genetic variation using a unique or various genetic operators (Ferreira, 2006). One of the advantages of GP (i.e. GEP) over other heuristic techniques (e.g. NF, NN and SVM) is in giving explicit expression of the input-target relationship. Gene Xpro program was used in the present study for GEP-based modeling. Different fitness functions and function sets were tried in the applied models and the best ones were selected. Details for model development will be given in the next sections. Further details about modeling process with GEP can be read in e.g. Ferreira (2006).

2.2. Neuro-fuzzy systems (NF)

Neuro-fuzzy technique (NF) is a combination of adaptive artificial neural network and fuzzy inference systems, where the parameters of the fuzzy system are computed by the neural networks training algorithms. NF calculates a set of parameters via a hybrid learning rule composed of back-propagation gradient descent error (BPGDE) and a least squared error (LSE). The Sugeno's fuzzy approach (Takagi and Sugeno, 1985) was used here to relate the target variable (*CEC*) to input variables. Different membership functions were evaluated here to find the optimal one. The hybrid optimization method (the combination of LSE and BPGDE) was used for obtaining the membership functions parameters to emulate the training data. For a given input-output matrix, various fuzzy-Sugeno models can be employed using different identification methods (i.e. grid partitioning and subtractive clustering, etc.). The commonly used grid partitioning (GP) identification method was utilized here for modeling soil CEC. The GP method proposes independent partitions of each antecedent variable by defining the membership functions (MFs) of all antecedent variables. Fuzzy MFs might take different forms and the optimal numbers of MFs is computed by trial and error. In selecting the number of MFs, large numbers of MFs or parameters should be avoided to save time and computational costs (Kisi and Shiri, 2012). For this reason, 2 or 3 numbers of MFs were used in the applied ANFIS models. Details for NF model structures used in the applications will be provided in the next sections.

2.3. Neural networks (NNs)

Neural networks are parallel information-processing systems which have been originally designed for the modeling of the performance of a biological neural system. The most common architecture of NNs is composed of the input, hidden, and output layers, which is called multilayer perceptron (MLP) (Fausset, 1994). Here, three-layer feed-forward networks were utilized with different transfer functions in the hidden and output layers. The hidden-layer-node numbers of each model were determined iteratively. At each training process, 100 networks were evaluated and the optimum architecture for each case (transfer functions) was selected. Also minimum and maximum values of 0.0001 and 0.001 were found to be optimum values of weight decay in hidden layer.

With modeling CEC through NN technique, the input and output values were normalized using the following equation:

$$h_{ni} = a\frac{CEC_i - CEC_{\min}}{CEC_{\max} - CEC_{\min}} + b \quad (1)$$

where CEC_{ni} is the normalized data at time i, CEC_{min} and CEC_{max} denote the minimum and maximum of the data set and CEC_i stands for the observed *CEC* value at time i. Different values can be assigned for the scaling factors a and b. The a and b were taken as 0.8 and 0.2 herein, respectively according to Cigizoglu (2003) and Kisi et al. (2013). Thus, the training data were scaled in the range [0.2, 0.8].

Detailed descriptions of NN techniques can be read in e.g. Bishop (1995) or Haykin (1999).

2.4. Support vector machine (SVM)

SVMs are regression procedures, with a structural risk minimization (SRM) principle formulation, which is superior to the traditional empirical risk minimization (ERM) principle, employed by conventional neural networks. Traditional ERM minimizes the error on the training data, while SRM minimizes an upper bound on the expected risk, providing SVM a greater ability to generalize, which is the goal in statistical learning (Vapnik et al., 1997; Gunn, 1998). Further details on the application of SVM can be found e.g. in Vapnik et al. (1997).

2.5. Gamma test for input selection

The Gamma test is a non-linear analysis and modeling method which allows examining the nature of a hypothetical input/output relationship in a numerical data-set. The Gamma statistic Γ is calculated utilizing the Gamma test and least Γ indicates the best input combination. First reported in Stefánsson et al. (1997) with the conjecture that a very simple method (the Gamma test) could be utilized to directly estimate from a given input/output data set the extent to which the data identified from an underlying smooth model, even though the model was unknown.

The set of input vectors in this study are values of Clay, Silt, Sand, OC and pH. The corresponding output is the soil *CEC*. We assume that the input vectors contain factors which are useful for influencing the output *CEC*. A second assumption is that the

underlying relationship of the system under investigation is of the following form:

$$CEC = f(\text{Clay},\ \text{Silt},\ \text{Sand},\ \text{OC},\ \text{pH}) + r \quad (2)$$

where f is a smooth function and r is a random variable that represents noise. The domain of possible models is restricted to the class of smooth functions which have bounded first partial derivatives. Γ is the estimate of that part of the output variance that cannot be accounted for by a smooth data model. Let $x_{N[i,k]}$ denotes the kth nearest neighbor in terms of Euclidean distance xi $(M \geq i \geq 1)$, $(p \geq k \geq 1)$ being p = number of nearest neighbours, we take $p = 10$. The main equations needed to calculate the Gamma statistic Γ are:

$$\delta_M(k) = \frac{1}{M}\sum_{i=1}^{M}\left|\mathbf{x}_{N[i,k]} - \mathbf{x}_i\right|^2 \quad (3)$$

where $|\cdot|$ is the Euclidean distance and

$$\gamma_M(k) = \frac{1}{2M}\sum_{i=1}^{M}\left|y_{N[i,k]} - y_i\right|^2 \quad (4)$$

Detailed rigorous mathematical explanation of these theoretical issues that conform the foundation of the Gamma test can be obtained in Evans (2002) and Evans and Jones (2004).

We performed the Gamma Test analysis to select the best combination of inputs with which to train the heuristic models, since it is critical to devise a systematic feature selection scheme that gives guidance on choosing the most representative features for estimating *CEC*. Taking into account that we consider 5 possible input variables then the number of possible combinations is $2^5 - 1 = 31$. Among these possible input combinations we would like to choose those having the lowest Γ statistic. First, we define the notion of a *Mask*. A mask of length 5 (because its length is the same as the number of input variables that we wish to consider), is a string of 0s and 1s with length 5. Each binary digit indicates one particular variable. A "1" indicates to include that specific variable for the input combination and "0" indicates not to include it. Hence, the mask: 11111 means that we include all the 5 possible attributes as inputs to the model. The concept of a mask helps us to obviously present any particular combination of attributes (Iturraran-Viveros, 2012). In Table 1, we have the list of the ten best combinations ordered from the lowest Γ statistic to the smallest (for data set 1).

2.6. Data set

Data from two separate sites were utilized in the present study. The first site is Mohr region which is located in Fars province, Southwest Iran (Fig. 1). This region is located between latitudes of 27°25′N to 27°59′N and longitudes of 53°05′E to 52°21′E with an area about 1900 km^2. The topographic elevation values of the study area range from 282 m a.s.l to 1780 m a.s.l, while the main topographic elevation takes values over 1031 m a.s.l.

The soil textures in this site are loam, sandy loam, clay loam and silty clay loam. The dominant soil types include Lithic Leptosols, Gypsic Regosols, Haplic Calsisols, Calcaric Cambisols and Calcaric Solonchaks, which cover mountains, hilland, plains and colluvial fans.

The main land uses practiced in the study area are pastures and irrigated farming across the Mehran River, characterized by arid climate with mean annual precipitation of 245 mm and mean annual temperature of 26.5 °C. A total 186 disturbed soil samples were obtained from two-first vertical depths (0–30 and 30–60 cm depth) of 93 representative soil profiles.

The second site is located in Mashhad Plain with an area of 6131 km^2, Khorasan-e-Razavi Province, Northeast Iran (Fig. 2). The region is located between 35°59′N to 37°04′N latitudes and 58°22′E to 60°07′E longitudes. The topographic elevation values of the study area range from 900 m a.s.l to 1500 m a.s.l, while the main topographic elevation ranges over 1200 m a.s.l. The soil textures are loam, sandy loam and sandy clay loam. The prevailing soil types include Calcaric Cambisols, Gypsic Regosols, Calcaric Regosols and Calcaric Fluvisols, which cover pediment plains, plateau and upper terraces and gravelly colluvial fans, respectively. The main land use in the study area is irrigated farming, in regions characterized by semi-arid climate with mean annual precipitation values of 222.1 mm and mean annual temperature of 15.8 °C (Keshavarzi et al., 2016). A Digital Elevation Model (DEM) with 10×10 m grid size was extracted from a paper-based topographic map using GIS platform with 1:25,000 scale and 10 m contour lines interval. As sampling is constrained by financial resources, accurate sampling strategies are desirable. Total 70 disturbed soil samples were obtained from two-first vertical depths (0–30 and 30–60 cm depth) of 35 representative soil profiles.

In both the sites, the sampling points were designed to cover evenly the whole area and to include different soil and land use types. Collected soil samples were air dried, crushed and sieved by using 2 mm sieve size. Large plant material and pebbles in each sample were separated by hand and discarded. Nonetheless, soil organic carbon (SOC) content was determined by the Walkley–Black method with dichromate extraction and titrimetric quantization (Nelson and Sommers, 1986). Percentages of clay (<0.002 mm), silt (0.002–0.05 mm), and sand (0.05–2 mm) particles were measured by means of the sieving and sedimentation method (Gee and Bauder, 1986). Soil reaction (pH) was measured in saturated paste extract using a digital pH-meter (Thomas, 1996) and the CEC was determined by method of Bower et al. (1952).

2.7. Data splitting

Usually, application of heuristic models involves a single data set assignment, where the models are trained using a part of available data (from input-target matrix) and tested by the rest data which has not been applied in training the models. Although this might be an easy and common way for constructing the heuristic models, all the available patterns cannot be tested, and the results strongly depend from the data assignment adopted. Therefore, a complete scanning of the data using a k-fold testing approach might be a promising approach for incorporation of all data in train-test stages (Shiri et al., 2014). Accordingly, in the present study, a k-fold testing data assignment technique was adopted to assessing the models accuracy for both the applied data sets. Each data set was divided into 10 equal parts and each part was reserved for testing the applied models at each modeling phase. So, 40 models (10 models per heuristic technique) were evaluated for each data set.

Table 1
List of the ten best combinations.

Gamma statistic (Γ)	Mask: Clay, Silt, Sand, OC, pH
0.00662	11011
0.01824	00110
0.02309	11111
0.03321	01111
0.03366	10111
0.03427	10011
0.03868	00111
0.03979	10110
0.04116	10010
0.04874	01010

Note: Clay (%), Silt (%), Sand (%), OC (%), pH.

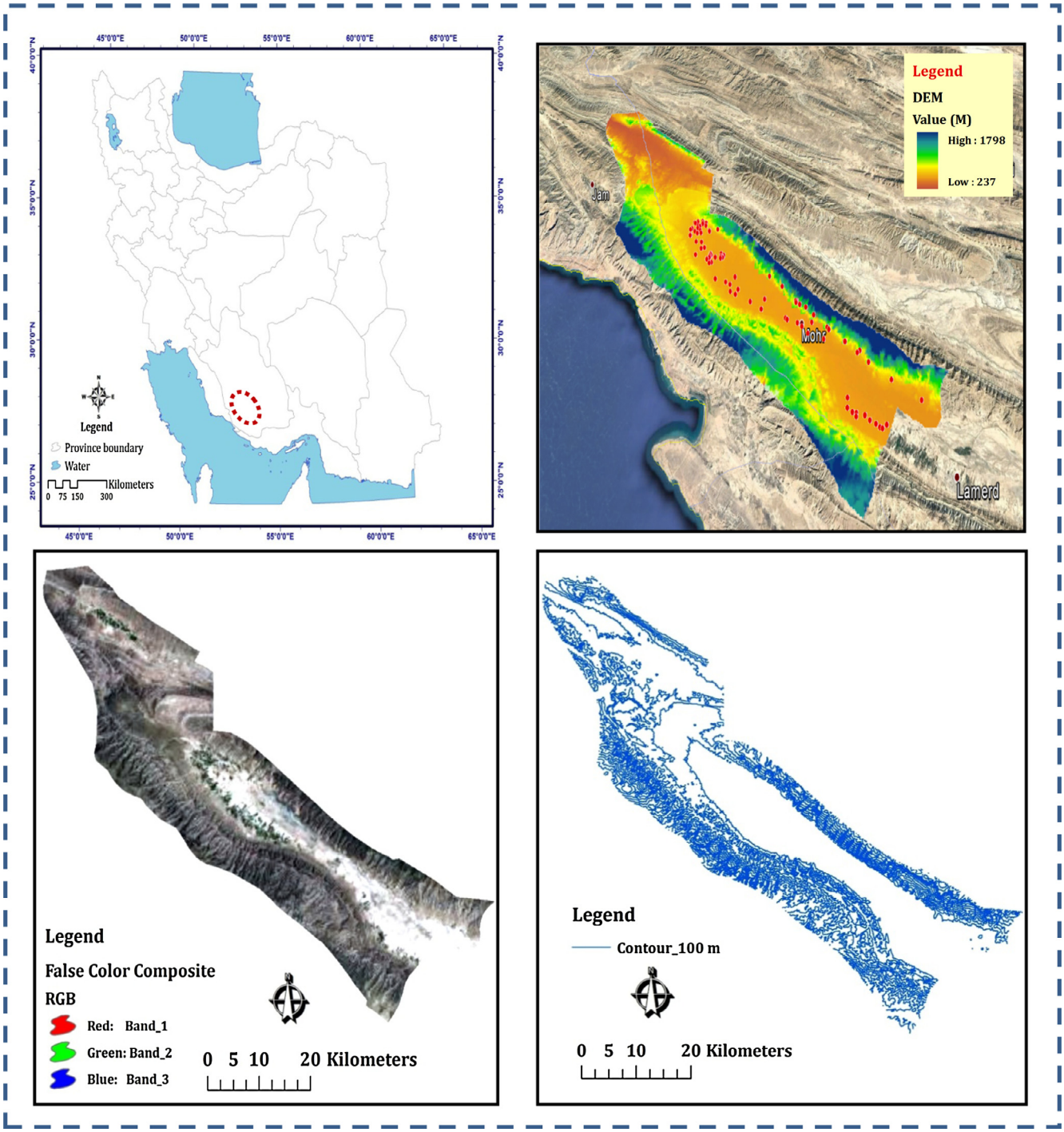

Fig. 1. Location of study area with sampling points, false color composite of google earth imagery and contour lines with 100 m interval (Mohr region, Iran).

2.7.1. Statistical criteria

Two statistical criteria were applied here for assessing the models performance, namely, the mean absolute error (*MAE*), and the scatter index (*SI*), expressions for which are:

$$MAE = \frac{\sum_{i=1}^{n} |CEC_{Oi} - CEC_{Mi}|}{n} \quad (5)$$

$$SI = \frac{RMSE}{meanCEC_O} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(CEC_{Oi} - CEC_{Mi})^2}}{meanCEC_O} \quad (6)$$

In these equations, CEC_O is the observed *CEC* value at the *i*-th observation step, CEC_M is the corresponding simulated *CEC* value, *n* is number of patterns, *mean* CEC_O is the mean value of the observations and RMSE is the root mean square errors. The indicators of the applied models were referred to the complete period for each data set, i.e. the simulations of each data set were pooled together and the statistical parameters were computed for the complete data set.

3. Results and discussions

As mentioned, the Gamma-test was applied here for identifying the models input variables. Consequently, the most influential

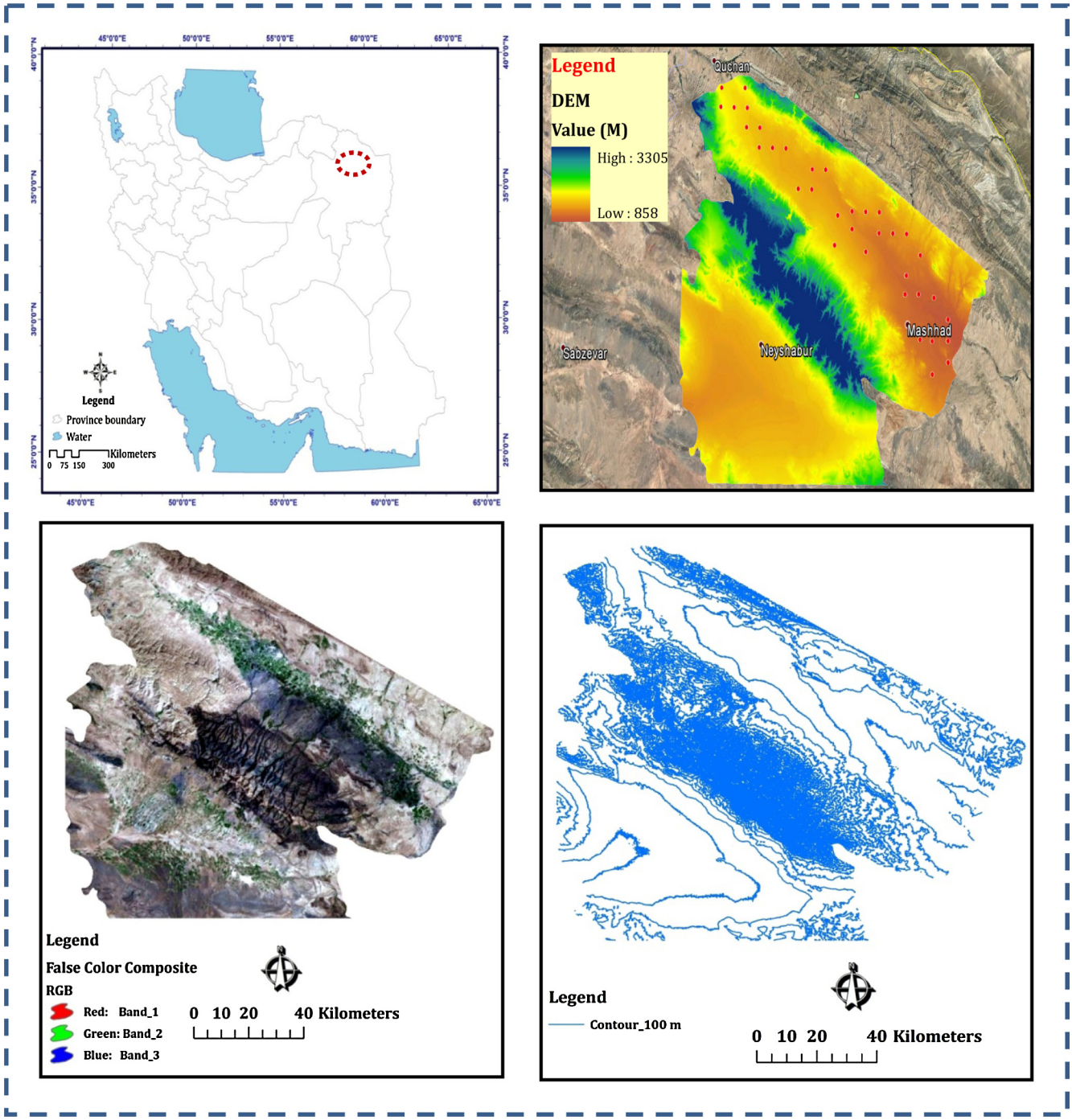

Fig. 2. Location of study area with sampling points, false color composite of google earth imagery and contour lines with 100 m interval (Mashhad plain, Iran).

parameters on soil CEC were determined for both the sites and utilized to feed the applied models.

3.1. Preliminary models structures

3.1.1. GEP parse tree

The first step with GEP modeling is to select the appropriate fitness function. For mathematical applications, small relative or absolute errors are usually utilized to discover an applicable solution (Ferreira, 2001a). The GEP model comprising all input parameters with default function set of GeneXpro (i.e., $+, -, \times, \div, \sqrt[3]{}, \sqrt{}, \ln, e^x, x^2, x^3, \sin x, \cos x, Arctgx$) was established for selecting the fitness functions. The results of this investigation are listed in Table 2. It is clear that the *RRSE* fitness function gives the most accurate results among others in both the locations. The next step consists of selecting the terminal and function sets. The terminal set includes {clay, silt, sand, OC, and pH}. For choosing the basic operators for creating the parse tree, different basic functions (which the learning algorithm explores them to derive a good model with independent variables) were investigated as listed in Table 3. A set of preliminary models were established to assess the models performance with using these function sets and selecting the best one. The results of the investigation are presented in Table 3 in terms of *SI*. From Table 3 it is seen that the function sets F4 and F3 surpass the

Table 2
Preliminary investigation on GEP fitness function using *SI*.

Fitness function based on the absolute error	*SI*	Fitness function based on the relative error	*SI*
Site 1			
Absolute error with selection range	0.044	Relative error with selection range	0.045
Absolute/hits	0.103	Relative/hits	0.104
Mean squared error (MSE)	0.044	r-MSE	0.047
Root mean squared error (RMSE)	0.043	r-RMSE	0.048
Mean absolute error (MAE)	0.044	r-MAE	0.045
Relative squared error (RSE)	0.043	r-RSE	0.046
Root relative squared error (RRSE)	0.042	r-RRSE	0.044
Relative absolute error (RAE)	0.046	r-RAE	0.045
Site 2			
Absolute error with selection range	0.130	Relative error with selection range	0.131
Absolute/hits	0.180	Relative/hits	0.188
Mean squared error (MSE)	0.126	r-MSE	0.133
Root mean squared error (RMSE)	0.124	r-RMSE	0.133
Mean absolute error (MAE)	0.131	r-MAE	0.132
Relative squared error (RSE)	0.127	r-RSE	0.128
Root relative squared error (RRSE)	0.120	r-RRSE	0.133
Relative absolute error (RAE)	0.130	r-RAE	0.132

Table 3
Preliminary investigation on GEP function set.

	Definition	Site 1	Site 2
F1	$\{+,-,\times,\div\}$	0.061	0.128
F2	$\{+,-,\times,\div,\sqrt{},x^2\}$	0.059	0.127
F3	$\{+,-,\times,\div,\sqrt{},Power,Lnx,Logx,e^x,10^x\}$	0.053	**0.117**
F4	$\{+,-,\times,\div,\sqrt[3]{},\sqrt{},\ln,e^x,x^2,x^3,Arctgx\}$	**0.042**	0.116
F5	$\{+,-,\times,\div,\sqrt[3]{},\sqrt{},\ln,e^x,x^2,x^3,\sin x,\cos x,Arctgx\}$	0.037	0.111
1	Addition	**0.042**	**0.117**
2	Multiplication	0.048	0.123
3	Subtraction	0.049	0.122
4	Division	0.058	0.128
5	Addition (with complexity increase)	0.069	0.134

Bold values represent the optimum models.

other four structures in locations 1 and 2, respectively. Although the F5 function set gives the most accurate results, the complexity of this function set for building parse tree is much higher than the others, as this set comprises different functions. So, this was omitted in developing GEP-based models. Table 3 also represents the sensitivity analysis of different GEP function settings. From the table it is clear that the addition linking function outperforms other linking options.

The next step consists of choosing the head length and genes numbers per chromosomes, which were selected as 8, and 3, respectively, as these numbers have been advised in literature (e.g. Shiri et al., 2014). Finally, the GEP operators should be selected, which were chosen as default values of GeneXpro, as advised by Shiri and Kisi (2011).

3.1.2. NF structure

Table 4 compares different NF structures in both sites. The second column of the table represents the final structures (number of membership functions) of NF models. In first line of the table, 3,2,3,2 indicate the number of MFs for the Clay, Silt, Sand, OC and pH inputs, respectively. As seen from Table 4, the NF model with triangular MFs provides better accuracy than the other models in both sites, which is in good agreement with the conclusions presented by Russel and Campbell (1996). Although the trapezoidal MFs have the same accuracy with triangular MFs, the latter is preferred because it has less number of parameters to be optimized.

3.1.3. NN structure

Four different NN models (Table 4) with different activation functions were employed at each site for modeling soil CEC: ANN1(4,5,1), ANN2(4,6,1), ANN3(4,3,1) and ANN4(4,7,1), for site 1, and ANN1(3,5,1), ANN2(3,6,1), ANN3(3,3,1) and ANN4(3,7,1) for site 2. Hidden neurons beyond these bonds didn't give accurate results. Here ANN1(4,5,1) indicates the neural network model comprising 4 input variables, 5 hidden nodes and 1 output node (CEC). The structures of each NN model are given in the second column of the table, representing the number of neurons at hidden layers. The number of hidden nodes was determined iteratively. NN models weights were adjusted using the conjugate gradient algorithm as advised by Kisi (2007). From Table 4, it is clear that the ANN3(4,3,1) and ANN3(3,9,1) having tansig activation functions in hidden and output neurons perform better than the other models in Site 1 and Site 2, respectively.

3.1.4. SVM structure

There are four main kernel functions used in SVM, namely, linear, sigmoid, polynomial and radial basis function. Table 4 represents the control parameters of each applied kernel function. The best fit kernel function is usually selected through sensitivity analysis. Therefore, this approach was employed here in modeling CEC. Table 4 sums up the *SI* values of different SVM models developed through using different kernel functions. The results clearly show that the radial basis function kernel outperforms the other kernels in both sites. The Gamma value of 0.25 was found to be the optimum kernel parameter for RBF by trial and error. Also number of 1000 iterations was applied and the error stop value was selected as 0.001.

3.2. Assessing the applied models

Table 5 sumps up the *MAE* and *SI* values of the GEP, NF, NN and SVM models for the both studied sites. As mentioned, the indicators of the applied models are referred to the complete period for each data set. The table clearly shows that the NF model gives the most accurate results in both sites with the lowest error values and the GEP, NN and SVM models can be ranked as the second, third and the fourth models, respectively.

Comparing the performance of the models between the studied sites, it is clear that the models give more accurate results in site 1 than those of the site 2, which might be linked to the data quality as well as the soil characteristics of these sites. NF model increased the accuracy of NN, GEP and SVM models by 4100%, 4000% and 6700% in predicting CEC in Site 1, respectively. Fig. 3 compares the measured and predicted CEC values by each method for Site 1. As observed from the scatterplots, the NF has the closest estimates to the corresponding measured CEC values while the SVM gives the worst estimates. As clearly seen from Fig. 4 where the estimates of the Site 2 are visually compared, the NF model has the least scattered estimates and its estimates closely follows the measured CEC values with higher R^2 value (0.929) than the other models. Here also the SVM has the worst estimates. Comparison of two figures (Figs. 3 and 4) shows that the models are much more successful in Site 1 than the Site 2.

Fig. 5 displays the *SI* values split up per test stage for both the studied sites. The individual accuracy of NN, GEP and SVM models per test stages fluctuated considerably more than for the NF models. GEP presents the maximum *SI* ranges between the best and worst values of 0.069 and 0.149 for Site 1 and Site 2, respectively,

Table 4
Statistics (*SI* values) of different ANFIS, ANN and SVM models.

	Structure		*SI*	
	Site 1	Site 2	Site 1	Site 2
NF models				
ANFIS1(Triangular MFs)	3,2,3,2	3,3,3	0.001	0.043
ANFIS2(Trapezoidal MFs)	3,3,3,3	3,3,3	0.001	0.065
ANFIS3(Generalized Bell MFs)	3,3,3,2	2,3,2	0.021	0.081
ANFIS4(Gaussian MFs)	3,3,3,3	3,2,3	0.002	0.052
ANFIS5(Two Gaussian MFs)	2,2,2,2	3,3,3	0.006	0.058
NN models				
ANN1(logsig, logsig)	4,5,1	3,5,1	0.043	0.017
ANN2(logsig, purelin)	4,6,1	3,5,1	0.050	0.020
ANN3(tansig, tansig)	4,3,1	3,9,1	0.041	0.015
ANN4(tansig, purelin)	4,7,1	3,6,1	0.048	0.018
SVM models				
SVM1 (linear kernel)	–	–	0.112	0.138
SVM2 (polynomial degree3 kernel)	0.11	0.11	0.104	0.134
SVM3 (sigmoid kernel)	0.11	0.14	0.098	0.129
SVM4 (radial basis function kernel)	0.50	0.18	0.068	0.125

Note: structures of the models in ANFIS, ANN and SVM, denote the number of membership functions, number of neurons, and control parameters (Gamma), respectively.

Table 5
Error statistics of the applied models.

	GEP	NF	NN	SVM
Site 1				
MAE	0.653	0.020	0.676	0.986
SI	0.042	0.001	0.041	0.068
Site 2				
MAE	1.446	0.412	1.357	1.495
SI	0.12	0.043	0.015	0.125

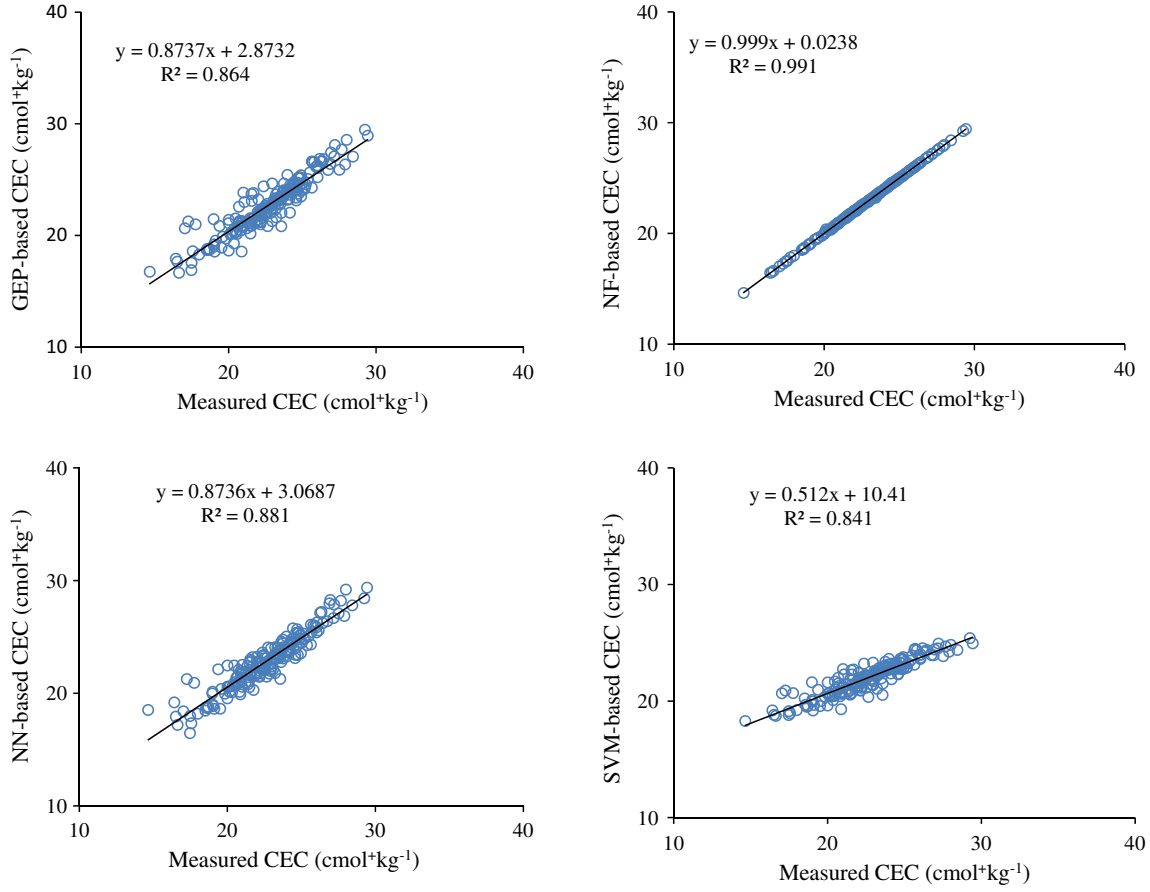

Fig. 3. Scatter plots of the observed vs. simulated CEC values for Site 1.

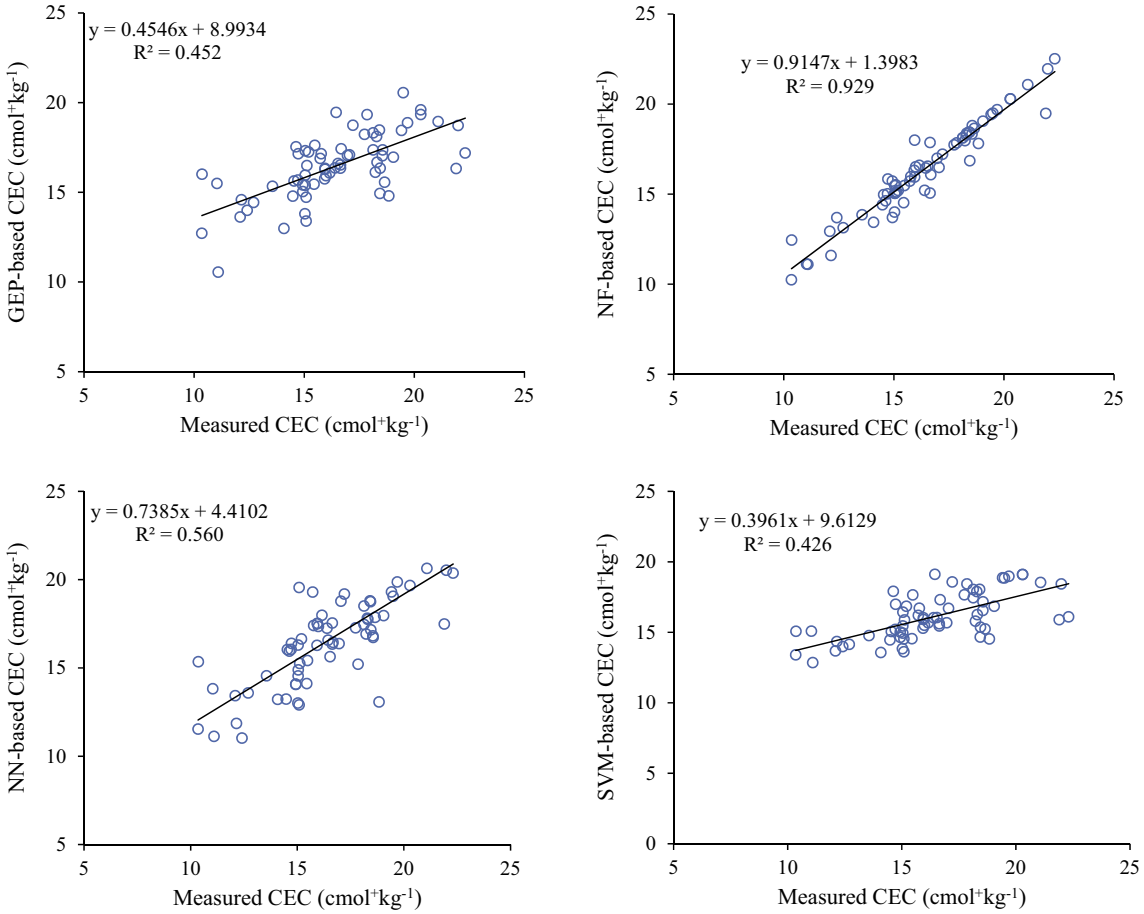

Fig. 4. Scatter plots of the observed vs. simulated CEC values for Site 2.

while, NF presents the minimum *SI* range values of 001 and 0.053 for the same sites.

With some exceptions, SVM approach presents the maximum *SI* values (minimum degree of accuracy) for both the sites. The performance variability within sites of GEP, NN and SVM models are considerably higher for NF. Although GEP and NN models showed similar average *SI* values for Site 1, Fig. 5 illustrates that the GEP model tends to be more accurate. Similar conclusions might be obtained for Site 2 where GEP and SVM presents the similar average *SI* values, while GEP seems to be more accurate as can be observed in Fig. 5. The performance fluctuations among test stages (data samples considered for testing) demonstrated the need to assess the models' performances through k-fold testing data set assignment procedures and not only considering a single data set assignment.

As mentioned, one of the advantages of the GEP models over other heuristic models is in providing the corresponding predictive mathematical expressions of the studied phenomenon. Table 6 presents the GEP expressions for both the sites as well as the contribution (weight) of each parameter on CEC simulation. Analyzing the expressions, it is seen that the GEP formulation for Site 1 (quadruple-input model) is more complicated than that of Site 2 (triple-input model). Both GEP expressions have much less parameters (much simpler) than the NF, NN and SVM models. For example, the optimal ANFIS1 model with 3,3,3 triangular MFs have $3 \times 3 \times 3$ (the number of MFs $\times$ the number of parameters in each MF $\times$ the number of inputs) premise parameters and 3^3 consequent parameters for Site 2. Similar to the ANFIS, the optimal ANN1(3,5,1) model have $3 \times 5 + 5 = 20$ parameters without bias terms.

OC and clay content have direct and significant effects on soil CEC and similar outcomes have been also reported in literature (Keshavarzi et al., 2017). It was also revealed that the effect of OC on CEC was significantly high, which might be due to high specific surface area and presence of functional group (Manrique et al., 1991). Similar results have been mentioned in literature, demonstrating the dominant effects of OC and/or soil clay content on CEC (Manrique et al., 1991; Keshavarzi et al., 2017).

The results were also tested using *t*-test for verifying the significance of differences between the observed and simulated CEC values. The statistics of the tests are given in Table 7. The NF model yields small testing values (−0.140 and −0.0003) with a high significance level (0.888 and 0.999) for the Site 1 and Site 2, respectively while the GEP model has slightly better statistics than the NF for the Site 1. According to the Table 7, the NF model seems to be more robust in Modeling CEC than the other models. The GEP models are also better than the NN and SVM models.

4. Conclusions

The present paper assesses the performance of generalized GEP, NF, NN and SVM techniques for soil *CEC* estimation using a k-fold testing data assignment approach. Accordingly, the test sets were independent of the training set, and complete data set is scanned for testing. Based on the obtained results, NF-based models are found to be more accurate than GEP, NN, and SVM-based models based on the availability of data. The NF method links the learning capability of the NN with the linguistic and transparent of a fuzzy logic. Hence, NF can be utilized for input/output mapping, similarly

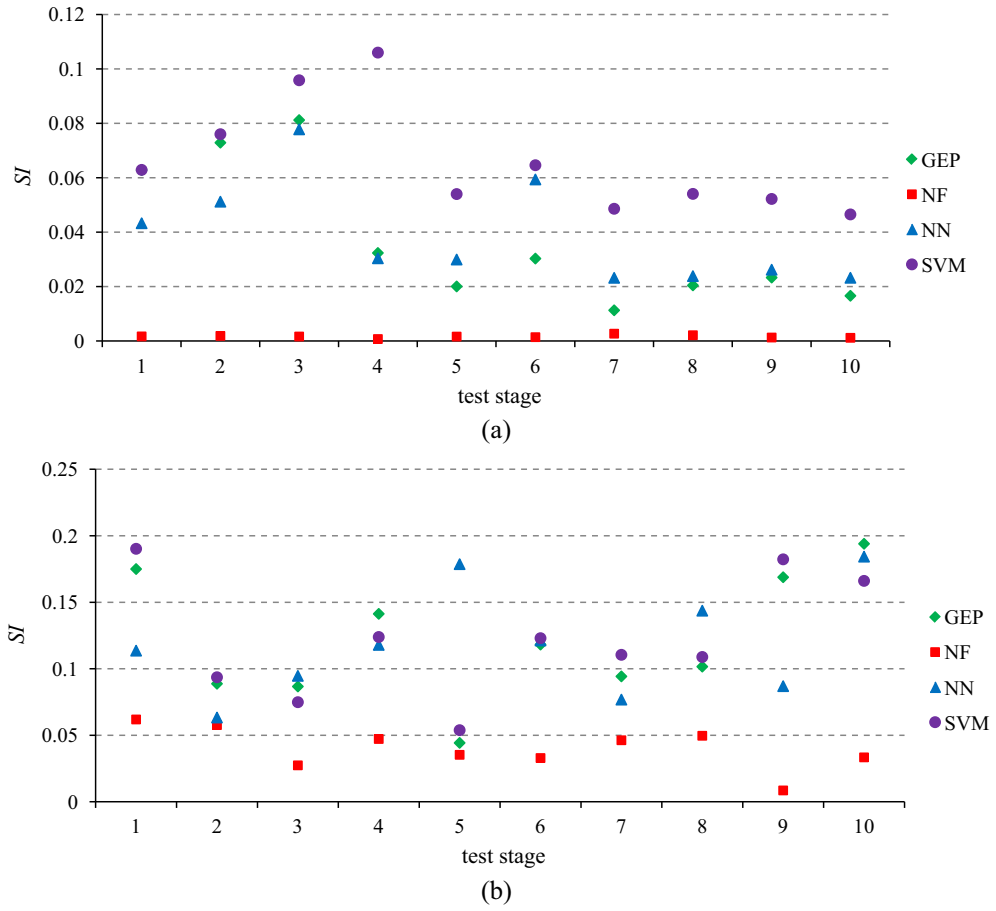

Fig. 5. *SI* values for CEC estimates split up per test stage: (a) Site 1, (b) Site 2.

Table 6
Mathematical expressions of the GEP predictive models.

Site	GEP expression	Parameters weights
Site 1	$CEC = \sqrt{Clay + Silt - OC^3 + 1.58532 Clay.OC^2} + Ln\left[pH^2(Silt - OC)\right] + 0.821\sqrt[3]{pH} + \left[\left(\frac{OC+pH}{Silt-2.5726}\right)^6 + OC\right]^3$	Clay = 2, Silt = 3 pH = 3, OC = 6
Site 2	$CEC = \log[2Clay - 13.969] + \sqrt{Clay - \sqrt{sand} + 2OC + 6.192} + Clay + 10.30$	Clay = 3, Sand = 1, OC = 3

Table 7
t-test of the applied models in modeling CEC.

	Site 1		Site 2	
	t-Statistic	Resultant significance level	*t*-Statistic	Resultant significance level
GEP	−0.125	0.900	−0.206	0.836
NF	−0.140	0.888	−0.0003	0.999
NN	−0.511	0.543	−0.527	0.599
SVM	1.512	0.198	1.148	0.254

as with an NN, however with the extra advantage of having the capacity to give the rules set on which the applied model is based. This provides additional intuition into the procedure being demonstrated. The performances of the applied models fluctuate throughout the test stages (within site) and between both sites. Therefore, a complete data set scanning would be necessary for suitable assessment of the applied models, since the conclusions based on a single data set assignment of the training and test sets might be misleading. The present paper used data from two sites with different soil characteristics. Further studies should be carried out using data from more stations with wide range of soil characteristics to reinforce these conclusions as well as for making an

external assessment of the models to generalize some heuristic models for the sites with data scarcity. In the present study, the predominant parameters affecting soil CEC were not similar in the studied sites, so not external modeling could be carried out. This might be a subject for the future studies.

Acknowledgements

This study was partially supported by Department of Soil Science, University of Tehran, Iran. Ursula Iturraran-Viveros acknowledges the support of Conacyt México under program PROINNOVA project number 231476. UI-V also thanks Pilar Ladrón de Guevara for her crucial help to locating useful references.

References

Bauer, A., Velde, B.D., 2014. Geochemistry at the Earth's Surface. Springer-Verlag, Berlin, Heidelberg.
Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press. 504pp.
Bishop, T.F.A., McBratney, A.B., 2001. A comparison of prediction methods for the creation of field extent soil property maps. Geoderma 103, 149–160.
Bower, C.A., Reitmeir, R.F., Fireman, M., 1952. Exchangeable cation analysis of saline and alkali soils. Soil Sci. 73, 251–261.
Cigizoglu, H.K., 2003. Estimation, forecasting and extrapolation of flow data by artificial neural networks. Hydrol. Sci. J. 48 (3), 349–361.
Emamgolizadeh, S., Bateni, S.M., Shahsavani, D., Ashrafi, T., Ghorbani, H., 2016. Estimation of soil cation exchange capacity using Genetic Expression Programming (GEP) and Multivariate Adaptive Regression Splines (MARS). J. Hydrol. 529, 1590–1600.
Evans, D., 2002. The Gamma test. Data derived estimates of noise for unknown smooth models using near neighbour asymptotics, Ph.D. thesis, University of Cardiff.
Fausset, L.V. (Ed.), 1994. Fundamentals of Neural Networks: Architectures, Algorithms and Applications. Prentice Hall, Upper Saddle River, NJ.
Ferreira, C., 2001a. Gene expression programming: a new adaptive algorithm for solving problems. Complex Syst. 13 (2), 87–129.
Ferreira, C., 2006. Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence. Springer, Berlin, Heidelberg, New York, p. 478.
Gee, G.W., Bauder, J.W., 1986. Particle size analysis. In: Klute, A. (Ed.), Methods of Soil Analysis: Part 1. Agronomy Handbook 9. American Society of Agronomy and Soil Science Society of America, Madison, WI, pp. 383–411.
Ghaemi, M., Astaraei, A.R., Sanaeinejad, S.H., Zare, H., 2013. Using satellite data for soil cation exchange capacity studies. Int. Agrophys. 27, 409–417.
Gunn, S.R., 1998. Support Vector Machines for Classification and Regression. Technical Report, University of Southampton, England.
Haykin, S., 1999. Neural Networks: a Comprehensive Foundation. Prentice-Hall, Upper Saddle River, New Jersey.
Hazelton, P.A., Murphy, B.W., 2007. Interpreting Soil Test Results: What Do All The Numbers Mean? CSIRO Publishing, Melbourne.
Iturraran-Viveros, U., 2012. Smooth regression to estimate effective porosity using seismic attributes. J. Appl. Geophys. 76, 1–12.
Johari, A., Habibagahi, G., Ghahramani, A., 2006. Prediction of soil–water characteristic curve using genetic programming. J. Geotech. Geoenviron. Eng. 132, 661–665.
Jones, A.J., 2004. New tools in non-linear modeling and prediction. Comput. Manage. Sci. 1, 109–149.
Keshavarzi, A., Sarmadian, F., Omran, S.W., Iqbal, M., 2015. A neural network model for estimating soil phosphorus using terrain analysis. Egypt. J. Rem. Sens. Space Sci. 18 (2), 127–135.
Keshavarzi, A., Omran, E.E., Bateni, S.M., Pradhan, B., Vasu, D., Bagherzadeh, A., 2016. Modeling of available soil phosphorus (ASP) using multi-objective group method of data handling. Model. Earth Syst. Environ. 2, 157.
Keshavarzi, A., Sarmadian, F., Shiri, J., Iqbal, M., Tirado-Corbalá, R., Ewis Omran, E., 2017. Application of ANFIS-based subtractive clustering algorithm in soil cation exchange capacity estimation using soil and remotely sensed data. Measurement 95, 173–180.
Kisi, O., 2007. Streamflow forecasting using different artificial neural network algorithms. ASCE J. Hydrol. Eng. 12 (5), 532–539.
Kisi, O., Shiri, J., 2012. River suspended sediment estimation by climatic variables implication: comparative study among soft computing techniques. Comput. Geosci. 43, 73–82.
Kisi, O., Shiri, J., Tombul, M., 2013. Modeling rainfall-runoff process using soft computing techniques. Comput. Geosci. 51, 108–117.
Koza, J.R., 1992. Genetic Programming: On the Programming of Computers by Means of Natural Selection. The MIT Press, Cambridge, MA, p. 840.
Makkeasorn, A., Chang, N.-B., Beaman, M., Wyatt, C., Slater, C., 2006. Soil moisture estimation in a semiarid watershed using RADARSAT-1 satellite imagery and genetic programming. Water Resour. Res. 42, 1–15.
Manrique, L.A., Jones, C.A., Dyke, P.T., 1991. Predicting cation exchange capacity from soil physical and chemical properties. Soil Sci. Soc. Am. J. 55, 787–794.
Nelson, D.W., Sommers, L.P., 1986. Total carbon, organic carbon and organic matter. In: Page, A.L. (Ed.), Methods of Soil Analysis: Part 2. Agronomy Handbook 9. American Society of Agronomy and Soil Science Society of America, Madison, WI, pp. 539–579.
Padarian, J., Minasny, B., McBratney, A., 2012. Using genetic programming to transform from Australian to USDA/FAO soil particle-size classification system. Austral. J. Soil Res. 50, 443–446.
Parasuraman, K., Elshorbagy, A., Si, B.C., 2007. Estimating saturated hydraulic conductivity using genetic programming. Soil Sci. Soc. Am. J. 71, 1676–1684.
Park, S.J., Vlek, L.G., 2002. Environmental correlation of three-dimensional soil spatial variability: a comparison of three adaptive techniques. Geoderma 109, 117–140.
Rayment, G.E., Higginson, F.R., 1992. Electrical Conductivity. In: 'Australian Laboratory Handbook of Soil and Water Chemical.
Rengasamy, P., Churchman, G.J., 1999. Cation exchange capacity, exchangeable cations and sodicity. In: Peverill, K.I., Sparrow, L.A., Reuter, D.J. (Eds.), Soil Analysis an Interpretation Manual. Melbourne, CSIRO.
Robertson, G.P., Sollins, P., Ellis, B.G., Lajtha, K., 1999. Exchangeable ions, pH, and cation exchange capacity. In: Robertson, G. Philip, Coleman, David C., Bledsoe, Caroline S., Sollins, Phillip (Eds.), Standard Soil Methods for Long-term Ecological Research. Oxford University Press, New York, NY, pp. 06–114.
Russel, S.O., Campbell, P.F., 1996. Reservoir operating rules with fuzzy programming. J. Water Resour. Plann. Manage. 122 (3), 165–170.
Shiri, J., Kisi, O., 2011. Comparison of genetic programming with neuro-fuzzy systems for predicting short-term water table depth fluctuations. Comput. Geosci. 37 (10), 1692–1701.
Shiri, J., Marti, P., Singh, V.P., 2014. Evaluation of gene expression programming approaches for estimating daily evaporation through spatial and temporal data scanning. Hydrol. Process. 28 (3), 1215–1225.
Stefánsson, A., Koncâr, N., Jones, A.J., 1997. A note on the Gamma test. Neural Comput. Appl. 5, 131–133.
Takagi, T., Sugeno, M., 1985. Fuzzy identification of systems and its application to modeling and control. IEEE Trans. Syst., Man Cybern. 15 (1), 116–132.
Thomas, G.W., 1996. Soil pH and soil acidity. In: Page, A.L. (Ed.), Methods of Soil Analysis: Part 2. Agronomy Handbook 9. American Society of Agronomy and Soil Science Society of America, Madison, WI, pp. 475–490.
Triantafilis, J., Lesch, S.M., Lau, L.K., Buchanan, S.M., 2011. Field level digital soil mapping of cation exchange capacity using electromagnetic induction and a hierarchical spatial regression model. Austral. J. Soil Res. 47, 651–663.
van Hoorn, J.W., van Alphen, J.G., 1994. Salinity Control. In: Ritzema, H.P. (Ed.), Drainage Principles and Applications. International Institute for Land Reclamation and Improvement, ILRI, The Netherlands, pp. 533–600.
Vapnik, V., Golwich, S., Smola, A.J., 1997. Support vector method for function approximation, regression estimation, and signal processing. In: Mozer, M., Jordan, M., Petsche, T. (Eds.), Advances in Neural Information Processing Systems 9, pp. 281–287.
Were, K., Bui, D.T., Dick, O.B., Singh, B.R., 2015. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecol. Indic. 52, 394–403.
Zolfaghari, A.A., Taghizadeh-Mehrjardi, R., Moshki, A.R., Malone, B.P., Weldeyohannes, A.O., Sarmadian, F., Yazdani, M.R., 2016. Using the nonparametric k-nearest neighbor approach for predicting cation exchange capacity. Geoderma 265, 111–119.