RESEARCH ARTICLE

Prediction of soil macronutrients using fractal parameters and artificial intelligence methods

Ali A. Zolfaghari (Zolfaghari, AA)[1], Meysam Abolkheiryan (Abolkheiryan, M)[1], Ali A. Soltani-Toularoud (Soltani-Toularoud, AA)[2], Ruhollah Taghizadeh-Mehrjardi (Taghizadeh-Mehrjardi, R)[3,4] and Amanuel O. Weldeyohannes (Weldeyohannes, AO)[5]

[1]Semnan University, Faculty of Desert Studies, Semnan, Iran [2]University of Mohaghegh Ardabili, Dept. of Soil Science and Engineering, Ardabil, Iran [3]Ardakan University, Faculty of Agriculture and Natural Resources, Ardakan, Iran [4]University of Tübingen, Dept. of Geosciences, Soil Science and Geomorphology, Tübingen, Germany [5]University of Alberta, Dept. of Renewable Resource, Alberta, Canada

Abstract

Aim of study: To evaluate artificial neural networks (ANN), and *k*-Nearest Neighbor (*k*-NN) to support vector regression (SVR) models for estimation of available soil nitrogen (N), phosphorous (P) and available potassium (K).

Area of study: Two separate agricultural sites in Semnan and Gorgan, in Semnan and Golestan provinces of Iran, respectively.

Material and methods: Complete data set of soil properties was used to evaluate the models' performance using a *k*-fold test data set scanning procedures. Soil property measures including clay, sand and silt content, soil organic carbon (SOC), electrical conductivity (EC), lime content as well as fractal dimension (D) were used for the prediction of soil macronutrients. A Gamma test was utilized for defining the optimum combination of the input variables.

Main results: The sensitivity analysis showed that OC, EC, and clay were the most significant variables in the prediction of soil macronutrients. The SVR model was more accurate compared to the ANN and *k*-NN models. N values were estimated more accurately than K and P nutrients, in all the applied models.

Research highlights: The accuracy of models among the test stages illustrated that using a single data set for investigation of model performance could be misleading. Therefore, the complete data set would be necessary for suitable evaluation of the model.

Additional keywords: artificial neural networks; Gamma test; *k*-Nearest Neighbor; support vector regression

Abbreviations used: ANN (artificial neural networks); CEC (cation exchange capacity); D (fractal dimension); EC (electrical conductivity); *k*-NN (*k*-Nearest Neighbor); MR (mean error); PSD (soil particle size distribution); R (correlation coefficient); RMSE (root mean square error); r-RMSE (relative root mean square error); SE (standard error); SOM (soil organic matter); SVR (support vector regression); ω (composite scaling constant)

Introduction

Soil nutrients play a considerable role in regulating plant growth. In soils with low nutrient content, limited productivity is expected from plants. On the other hand, excessive use of fertilizers may have detrimental consequences on environmental pollution, in the case they enter into the surface or the groundwater bodies (Wang

et al., 2009; Fu *et al.,* 2010). The yield of agricultural crops is completely dependent on the application of soil fertilizers, especially nitrogen (N), phosphorus (P), and potassium (K) (Foth, 1982). Ayoubi & Sahrawat (2011) illustrated that the soil total N and available P are the most important factors in predicting barley yield. Soil properties, including clay content, lime content as well as pH are negatively correlated with the mineral nutrients and may decrease the availability of plant nutrients by fixing them or by forming insoluble compounds (Chaudhari *et al.,* 2012). Additionally, electrical conductivity (EC) has significant influence on crop yield, crop suitability, activity of soil microorganisms, and availability of plant nutrients, which can consequently influence the key soil processes (Adviento-Borbe *et al.,* 2006).

The most accurate way for direct measurement of soil macronutrients is laboratory analysis. However, this analysis is difficult in arid and semi-arid regions of Iran, which have with calcareous soils. Alternatively, during the past three decades, data-driven models (*e.g.* artificial neural networks (ANN), support vector regression (SVR) and *k*-Nearest Neighbor (*k*-NN)) have emerged as rapid mathematical methods, which correlate the role of independent variables to dependent ones (Simon, 1995; Moosavi & Sepaskhah, 2012; Shiri *et al.,* 2017). A limited number of approaches have been proposed for the prediction of soil macronutrients and soil chemical properties utilizing data driven models (Li *et al.,* 2013; Keshavarzi *et al.,* 2015; Jeong *et al.,* 2017). For instance, Jeong *et al.* (2017) combined different vegetation indices with some attributes derived from digital elevation map in order to predict the spatial soil nutrients. Their results indicated that SVR is an efficient method for prediction of N and P nutrients. Li *et al.* (2016) investigated the capability of ANN model merged with kriging, in order to predict relationships among soil properties and environmental factors. Keshavarzi *et al.* (2015) found that the values of soil P could be predicted more accurately by the ANN-based pedotransfer function and by including topographic variables in the model. Safaa & Maxwell (2015) predicted pasture N content using ANN models and thermal images. Their results indicated that ANN models could be well fitted for pasture N content. Zolfaghari *et al.* (2016) applied the *k*-NN approach to obtain cation exchange capacity (CEC) using soil particle size distribution (PSD) and soil organic matter (SOM). Their results illustrated that the ANN and *k*-NN approaches have similar accuracies in estimating soil CEC.

In most reported literature on developing and testing data mining models, a single random data set has been utilized, and subsequently, using a portion of the whole data, models have been trained and been tested using the remained data pattern. In some cases, the efficiency of the models with suitable results, drastically decreases by substitution of the test data. Hence, it is more suitable to determine the efficiency of the models based on the *k*-fold testing method (Shiri *et al.,* 2017). In this study, we have investigated the capability of the ANN, SVR, and *k*-NN models in predicting the availability of soil macronutrients, using the *k*-fold procedure. In addition, the sensitivity analysis of the input variables has been determined using Gamma-test technique. To predict the availability of soil macronutrients, fractal parameters derived from the PSD were also used as the model input variables.

Material and methods

Study area and soil analysis

This study was conducted in two separate agricultural sites in Semnan and Gorgan, respectively from Semnan and Golestan provinces of Iran. Semnan region, in the Semnan province, is located in the central part of Iran. Semnan region has an arid climate with high evapotranspiration and very little precipitation. Soil samples in this area were selected from rangelands and agricultural lands. Rangelands contain poor vegetation cover, while 90% of agricultural lands are used for the production of wheat, barley and corn (Babaei *et al.,* 2018). The mean annual temperature, precipitation and potential evapotranspiration in the Semnan region are 18.5 ºC, 138 mm and 2500 mm yr^{-1}, respectively. Based on the FAO World Reference Base for Soil Resources (WRB) most soils of Semnan are classified as Regosols, Calcisols and Cambisols.

Gorgan, in the Golestan province, is located in the northern part of Iran. The Gorgan region has a humid climate. Soil samples on this area were selected from agricultural lands which were used for the production of wheat, barley, rapeseed and watermelon. The mean annual temperature, precipitation and potential evapotranspiration in the Gorgan region are 17.5 ºC, 550 mm and 985 mm yr^{-1}, respectively. Based on the WRB, most soils of this area are classified as Luvisols, Leptosols and Cambisols.

This work was conducted at five steps (Fig. 1): i) soil properties including pH, EC, OC, CaCO$_3$, clay, silt and sand were measured; ii) step fractal parameters including D and ω from PSD were determined; iii) the Gamma test was utilized for identifying the relevant variables for modeling of soil macronutrients; iv) soil macronutrients were predicted using ANN, SVR and *k*-NN models; v) the best model was selected based on the cross validation.

For determining the relation of soil macronutrients (*i.e.* soil N, available P and available K) with easily measurable soil properties, 130 soil samples from the top 30 cm of soil profile were collected from two separate agricultural
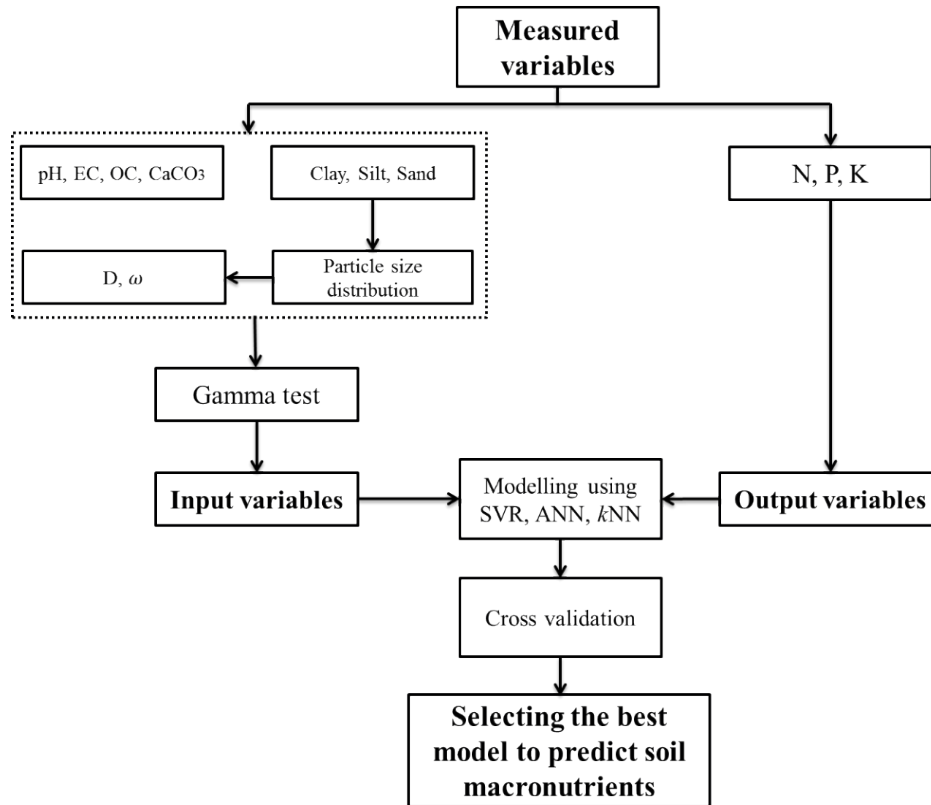
Figure 1. Flowchart of the procedure summarizing the sequence of methodologies applied for prediction of macronutrients. EC: electrical conductivity; OC: organic carbon; D: fractal dimension of the particle size distribution (PSD); ω: composite scaling constant; SVR: support vector regression; ANN: artificial neural network; k-NN: k-nearest neighbors.

sites in Semnan and Gorgan, respectively from Semnan and Golestan provinces of Iran. The sampling points were determined using a random sampling scheme.

Clay (<0.002 mm), silt (0.002–0.05 mm), and sand (0.05–2 mm) particles were measured using sieving and sedimentation technique (Gee & Bauder, 1986). Soil organic carbon was defined by Walkley–Black approach (Walkley & Black, 1934), and soil reaction (pH) and electrical conductivity (EC$_{25}$ °C) were obtained from extract paste (Jackson, 2005). N, P and K nutrients were respectively measured by a modified Kjeldahl wet digestion procedure and a Tector Kjeltec System, the Olsen method (Olsen & Khasawneh, 1980) and the flame photometry (Jackson, 2005).

Fractal parameters

Fractal parameters, estimated from the PSD, were also used as the model input variables for prediction of the desired studied macronutrients. For calculating fractal parameters, first, the PSD curve was developed from limited soil texture data using Skaggs *et al.* (2001) method. At the second step, Bird *et al.* (2000) fractal model

(equation 1) was fitted on the PSD curve data and fractal parameters were determined. The PSD for the 20 size classes (0-2, 2-3, 3-5, 5-10, 10-20, 20-30, 30-40, 40-50, 50-60, 60-100, 100-150, 150-200, 200-300, 300-400, 400-600, 600-800, 800-1000, 1000-1300, 1300-1600 and 1600-2000 μm) were calculated from sand, silt and clay, applying the model described by Skaggs *et al.* (2001) as follows:

$$P_r(r) = \frac{1}{\left(1 + \frac{1}{P_r(r_0)} - 1\right) exp(-ur^\sigma)} \quad (1)$$

where $P_r(r)$ is the mass fraction of soil particles with a radius < r; r_0 is the lower limit on radius for which the model applies, and σ and u are the model parameters which can be estimated as follows:

$$\sigma = \alpha ln \frac{v}{\lambda} \quad , \quad u = -v^{1-\beta}\lambda^\beta \quad (2)$$

$$v = ln \frac{\frac{1}{P_r(r_1)} - 1}{\frac{1}{P_r(r_0)} - 1} \quad , \quad \lambda = ln \frac{\frac{1}{P_r(r_2)} - 1}{\frac{1}{P_r(r_0)} - 1} \quad (3)$$

$$\alpha = \frac{1}{ln \frac{r_1 - r_0}{r_2 - r_0}} \quad , \quad \beta = \alpha ln \frac{r_1 - r_0}{r_0} \quad (4)$$

$$1 > P_r(r_2) > P_r(r_1) > P_r(r_0) > 0 \, , \quad r_2 > r_1 > r_0 > 0 \quad (5)$$

To predict PSD, the values of r_0, r_1, and r_2 must be predefined. In this study, the values of r_0, r_1, and r_2 were selected to be equal to 1, 25, and 999 μm, respectively (according to Skaggs *et al.,* 2001). Considering the USDA particle-size classification system, these radii indicate that P_r (r_0) is the clay mass fraction, P_r (r_1) is clay plus silt mass fraction and P_r (r_2) is the clay plus silt plus fine sand plus very fine sand mass fraction (Skaggs *et al.,* 2001; Bayat *et al.,* 2014).

In the second step the Pore-Solid Fractal model was fitted on PSD data (Bird *et al.,* 2000):

$$M_s(r \leq r_i) = \omega r_i^{3-D} \qquad (6)$$

where $M_s(r<r_i)$ is the cumulative mass of particles smaller than r_i, D is the fractal dimension of the PSD, and ω is the composite scaling constant.

Defining the optimal combination with the Gamma test

The Gamma test is an appropriate tool for identifying the optimum combination of the input variables (Yazdani & Zolfaghari, 2017). In this research, the Gamma test was applied in order to find the most effective parameters. Then the least sensitive parameters were excluded from the models. We considered the following set:

$$\{(x_i, Y_i), 1 \leq i \leq M\} \qquad (7)$$

where $x_i W R^m$ are considered as m dimensional input vectors (with a record length of M) with real numbers (R) and, accordingly, $Y_i \epsilon R$ are output scalper. In addition, it is assumed that x vectors have the capability of predicting factors affecting the output Y.

The following equation explains the relationship between x and Y vectors.

$$Y = f(x_1 \dots x_m) + e \qquad (8)$$

where f refers to a smooth function and e shows a random noise with zero mean.

The Gamma statistic (Γ) refers to an estimate of output variance in the model. The Gamma test is defined by NE[i,k], that are the kth ($1 \leq k \leq n$) nearest neighbors. Specifically, the Gamma test is derived from the Delta function of the input vectors:

$$\delta_M(k) = \frac{1}{M} \sum_{i=1}^{M} |x_{NE[i,k]} - x_i|^2 \quad (1 \leq k \leq n) \qquad (9)$$

where $|\dots|$ refers to the Euclidean distance. Subsequently, the corresponding Gamma function of the output values obtains as:

$$\gamma_M(k) = \frac{1}{2M} \sum_{i=1}^{M} |Y_{NE[i,k]} - Y_i|^2 \quad (1 \leq k \leq n) \qquad (10)$$

where $Y_{NE[i,k]}$ is the corresponding Y-value for the kth nearest neighbor of x_i in Eq. (9). For calculating Γ, a least-square regression line is determined for the n points $(\delta_M(k), \gamma_M(k))$

$$\gamma = A\delta + \Gamma \qquad (11)$$

where A indicates the slope of linear regression. Intercept of the above regression line is equal to Γ. A higher slope indicates a higher level of complexity in the model.

The Γ statistics can be standardized for the creation of the V ratio. The V ratio was computed as follows:

$$V - ratio = \frac{\Gamma}{\sigma^2(Y)} \qquad (12)$$

where σ^2 (Y) refers to variance of the output variable. Lower V ratio indicates a better prediction of the output variable, while higher V represents a greater prediction error (Moghaddamnia *et al.,* 2009; Yazdani & Zolfaghari, 2017).

Standard error (SE) of the regression give us an indication of how much the point estimate is likely to vary from the corresponding data. It can be calculated as square root of the estimated error variance on the sum of square of the independent variable. Therefore, if the SE value of linear regression in Eq. (9) is close to zero, we have more confidence in the value of the Gamma statistic as an estimate for the noise variance on the given output.

Prediction models

Artificial neural network (ANN)

ANN include the input, output and the hidden layers we used. Mathematically, it is expressed by:

$$O_k = S\left(\sum_{j=1}^{m} W_{jk} \times S\left(\sum_{i=1}^{n} W_{ij} X_i\right)\right) \qquad (13)$$

where O_k and X_i are the output and input variables of the neural network, W_{ij} and W_{jk} are the weights between the input and hidden layers, and the weights between the hidden and output layers, respectively. S refers to a transfer function, m and n, the numbers of input and hidden neurons, respectively (Haykin, 1994). The number of hidden layers represents the complexity of the ANN. It is certified that various soil properties could be successfully predicted using one hidden layer in ANN (Were *et al.,* 2015; Taghizadeh-Mehrjardi *et al.,* 2016). Therefore, in the current study a topology with one hidden layer was utilized. There are various methods for optimizing the number of neurons inside the hidden layer. For obtaining the best neuron number in the hidden layer, ANNs with 2 to 20 neurons in the hidden layer were compared. The best estimated neuron number in the hidden layer was chosen by a cross-validation using a 5-fold process over the training set. In this study, we applied the sigmoid function as the transfer function for prediction of soil properties. ANN can be trained by a back propagation algorithm. Among various back-propagation training algorithms, we

selected the Levenberg–Marquardt due to its efficiency and simplicity. Neural Network Toolbox of MATLAB was utilized for performing ANN modeling (Mathworks, 2010).

k-*nearest neighbors (*k-*NN)*

The *k*-NN algorithm proposed by Nemes *et al.* (2006), Botula *et al.* (2013) and Zolfaghari *et al.* (2016) was used to predict soil macronutrients. MATLAB environment was utilized for implementing the *k*-NN algorithm (Mathworks, 2011). This approach does not need any predefined mathematical function for estimating the target variable. In the *k*-NN technique, soils which have the most similarity to the target soil (test sample) are chosen from a reference data set. The efficiency of the *k*-NN technique is greatly dependent on the correct selection of the 'most similar' soils. The details of the *k*-NN algorithm are:
— Step 1: Defining the test (target) and the reference data set.
— Step 2: Calculating the Euclidean distance between test and training samples. The similarity between the target (test data set) and the reference soils obtains using the Euclidean distance was calculated:

$$d_i = \sqrt{\Sigma_{j=1}^{np} \Delta a_{ij}^{\,2}} \qquad (14)$$

where d_i refers to the "Euclidean distance" of the ith soil from the target soil and Δa_{ij} indicates the difference among the ith soil and the target soil in the jth soil attribute and *np* shows the number of soil properties considered for the model. The term 'distance' determines the similarity; the distance will be smaller for soils which have higher similarity with the target soil regarding the input attributes.
— Step 3: Determining *k* nearest neighbors by sorting the distances in ascending order, and then choosing *k* samples with relative minimum distances.

The normalization procedure is necessary because of significant differences among soil variables. For this reason, the input variables are first transformed to temporary attributes with a distribution having zero mean and one standard deviation.

Finally, the reference soils data set are sorted in ascending order and according to their distance from the target soil. The distances of the selected *k* neighbors from the reference soil can be obtained using a weighting procedure as below:

$$w_i = \frac{d_{i(rel)}}{\sum\limits_{i=1}^{k} d_{i(rel)}} \qquad (15)$$

where *k* refers to the number of selected neighbors, w_i shows the weight associated with the ith nearest neighbor, and *di*(*rel*) defines the relative distance of the ith selected neighbor, calculated as:

$$d_{i(rel)} = \left(\frac{\sum\limits_{i=1}^{k} d_i}{d_i}\right)^p \qquad (16)$$

where *p* is a power term which considers different possible weight–distance relationships.

After calculation of w_i, soil macronutrients would be predicted as:

$$\hat{Y}_i = \sum_{i=1}^{k} w_i Y_i \qquad (17)$$

where Y_i, $\hat{Y}_i$ are the observed and the predicted soil macronutrients for i^{th} soil sample.

Support vector regression (SVR)

SVR is a strong data mining approach that can be generalized to non-linear models applying a kernel function (Vapnik, 1995). The kernel function is a set of mathematical functions, which is used to transform the input data space into a higher dimensional space. In this study ε-SVR was used for the prediction of soil macronutrient (Smola & Schölkopf, 2004). In ε-SVR a threshold (ε) set is given by the user. Data points with an absolute difference greater than the threshold, contribute to the regression fit. Since the squared residuals are not used in ε-SVR, large outliers have limited effects on the regression equation. Also samples which fit suitably on the model (*i.e.,* with small residuals) have no effect on the regression equation. The best SVR fit in the features space utilizes a loss function in order to find the regression line that minimizes the model residual. The loss function in ε-SVR applies absolute residuals over the threshold ε; however it ignores all errors within ± ε. This approach shows robust effects that are not sensitive to outliers on the prediction (Cherkassky & Mulier, 2007). With ε, the loss function can add a penalty (cost) for large residuals. There is a relationship between the ε and the cost parameter (Kuhn & Johnson, 2013). At the initial step, 0.1 (default value) was chosen for ε, and kept fixed while tuning the value of cost via 10-fold cross validation. In this study, the radial basis kernel function (Eq 18) performed the greatest efficiency because its parameters were easier to adjust compared to those of other kernels (Ballabio, 2009; Kuhn & Johnson, 2013).

$$G(x_i, x_j) = exp\left(-\sigma \|x_i - x_j\|^2\right) \qquad (18)$$

Table 1. Basic statistics of input and output soil variables

	Input variables									Output variables		
	Clay	Silt	Sand	pH	EC (dS/m)	OC	CaCO₃	D	ω	N	P	K
	(%)					(%)				(mg/kg)		
Minimum	1.32	2	10.56	5.95	0.34	0.01	1	2.28	0.003	0. 06	0.03	110.44
Maximum	51	85	93	8	60	7	52	2.98	0.84	9.6	29.50	1452
Average	19.4	37.1	43.5	7.5	3.2	1.5	18.7	2.88	0.44	2.0	6.7	605.8
Std. deviation	12.56	22.7	20.08	0.53	7.48	1.59	13.34	0.10	0.21	1.90	8.11	287.58
Median	17.24	30	41.38	7.67	1.16	1.10	16.75	2.91	0.44	1.0	3.51	609.85
Skewness	0.41	0.78	0.45	-1.20	6.42	2.02	0.61	-2.74	-0.03	1.78	2.14	0.55
Kurtosis	-1.12	-0.4	-0.09	0.6	45.34	4.29	-0.32	9.54	-0.44	6.30	5.91	0.11

EC: electrical conductivity; OC: organic carbon; D: fractal dimension of the particle size distribution (PSD); ω: composite scaling constant; N: nitrogen; P: phosphorus; K: potassium

where G represents the kernel function, x indicates the input vectors, σ refers the scaling parameter, and $\|\ \|$ shows the Euclidean norm. The radial basis function needs to adjust the sigma (σ) parameter that controls its width. In this research, σ was automatically optimized in MATLAB software by minimizing *k*-fold cross-validation.

Data splitting and model assessment

A *k*-fold testing data assignment approach was applied for the analysis of efficiency of the studied models, namely *k*-NN, ANN and SVR, which were used for the prediction of soil macronutrients. In this step, the variables significant for prediction of each nutrient (*e.g.* N, K, and P), determined using the Gamma test, were used. The complete data was partitioned into 10 subsets. Each time, the models were trained and were tested utilizing a part of the available patterns. Using this approach, all the available input-target patterns were considered by the models to construct the final estimation model (Marti *et al.,* 2013; Shiri *et al.,* 2014, 2017). Accordingly, the *k*-NN, ANN and SVR models were trained and tested 30 times (3 approaches 10-folds). Assessing the accuracy of the models' performance using the *k*-fold testing would circumvent from acquiring relatively rational conclusions that might be drawn using traditional data management scenarios (Marti *et al.,* 2013; Shiri *et al.,* 2014). This was possible because no invisible input patterns would remain in the development of the models. The performance accuracy of the applied models was evaluated based on four statistical metrics: correlation coefficient (R), root mean square error (RMSE), mean error (MR) and relative root mean square error (r-RMSE). MR illustrates systematic errors in the applied approach, RMSE represents the accuracy of the approach; r-RMSE is a dimensionless indicator that can provide a proper and consistent contrast among the models when they depend on different targets. This is extremely decisive as the influence of the targets'

magnitude is neglected because in calculation of r-RMSE, only their mean values are taken into account. These metrics can be obtained by:

$$R = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)\left(\hat{Y}_i - \overline{\hat{Y}}\right) \Bigg/ \sqrt{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{\hat{Y}}\right)^2} \tag{19}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)^2} \tag{20}$$

$$MR = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right) \tag{21}$$

$$r - RMSE = \frac{RMSE}{\overline{Y}} \tag{22}$$

where *n* represents the number of data; Y_i and $\hat{Y}_i$ refer to the observed and the predicted soil macronutrients for the *i*th soil sample, respectively; and the $\overline{Y}$ and $\overline{\hat{Y}}$ are the mean observed and predicted soil macronutrients, respectively.

Results and discussion

Data summary statistics

Relevant statistics of input and output soil properties of the existing and the newly evolved models are represented in Table 1. The data sets embed a wide range of soil factors. Among the input variables, soil PSD covers a good range and can be expandable to many regions with different climatic conditions. Soil pH changed from 5.95 to 8, with an average of 7.5. Soil EC varied from 0.34 to 60 dS/m in the studied regions, which indicates a wide and expandable range (Table 1). Moreover, organic carbon had acceptable fluctuation ranging from 0.01 to 7%. The values of macronutrients showed considerable fluctuation. Nitrogen content had negligible values in poor and infertile soils but showed noticeable values in fertile soils. Variation of soil P was higher than the other macronutrients but its average indicated P deficiency in most soils.

The texture of the parent material showed high K content. Large variation in soils macronutrients was largely due to the use of chemical, manures, and mismanagement.

Sensitivity analysis

The Gamma value was calculated for a combination of nine input variables listed in Table 1. Recognizing their role, the input variables were omitted one by one from the initial combination (Table 2). In addition to the Gamma values, V ratio and the SE were also calculated. Increase of the Gamma value after omitting a variable means that the omitted variable has a profound positive effect on the model, and therefore it should be included in the initial combination. Conversely, decrease of the Gamma value after omitting a variable, means that the omitted variable has a weak influence on the model and should be excluded from the initial combination.

Based on the results, Gamma values for N, P and K in the first combination were 0.0002, 0.081 and 0.084, respectively. Calculated Gamma values from other combinations were compared with the preceded numbers and subsequently the effective parameters of the model were defined. The results showed that for N, five variables have positive effects on Gamma value. The highest effect was observed from the OC, EC, silt content, lime content and ω (Table 2). Kovačević *et al.* (2010) and Jeong *et al.* (2017) reported a strong positive correlation between C and N. For P, four variables (pH, OC, clay and fractal dimension (D)), had significant influence on Gamma value. The most important variables for predicting K were pH, EC, OC, clay and lime content. On the other hand, the results indicated that D had the lowest effect for the prediction of N (Gamma value = 0.0005) while for P and K the lowest effects were related to sand and silt, respectively. The results illustrated the importance of clay content as an effective variable for the prediction of K and P. These findings are reasonable because clay minerals transformation is strongly dependent on K amounts of soil and also the dynamics causing either release or removal of K to or from the soil (Raheb & Heidari, 2012). Raheb & Heidari (2012) reported a strong correlation between K and clay content in the soil. Moreover, EC and OC were evaluated as the most important factors in predicting soil macronutrients. In addition, Gamma values between CaCo3 and N and K values were noticeable. Decomposition of SOM increases ions in soil liquid phase. Such modifications have negative impact on soil EC, which itself is tuned by several soil fertility variables like K, Ca, Mg, pH, N, P, CaCO3, CEC, and OC (Bronson *et al.,* 2005; Aimrun *et al.,* 2009; Peralta & Costa, 2013).

Sensitivity analysis showed that the soil pH was an important variable in the prediction of K since most of the studied soils were slightly alkaline (mean and median of pH in the studied areas were 7.5 and 7.6, respectively). By increasing pH in soils, K enters the hexagonal cavities. Consequently, K fixation increases and it becomes inaccessible. The negative relation between pH and K obtained by the regression analysis proves this hypothesis.

A higher V-ratio also suggests the importance of combination for modeling. This V-ratio corresponds well to the resulting Gamma value (Fig. 2). Therefore, a higher Gamma value corresponds to a higher V-ratio. The greatest and the least V-ratios in the prediction of N were related to OC (0.376) and D, respectively. The highest V-ratio in the prediction of P and K was related to sand (0.251) and silt (0.317), respectively. On the other hand, the lowest V-ratio in the prediction of P and K was related to soil pH (0.50, 0.42). These results explain that soil pH

Table 2. Comparison of the Gamma values in different soils macronutrients

Combination	Omitted variable	N		P		K	
		Gamma value (Γ)	Standard error	Gamma value (Γ	Standard error	Gamma value (Γ)	Standard error
1	-	0.0002	0.009	0.081	0.031	0.084	0.017
2	EC	**0.008**	0.009	**0.104**	0.032	**0.103**	0.016
3	pH	0.004	0.009	**0.125**	0.021	**0.105**	0.015
4	OC	**0.094**	0.020	**0.102**	0.040	**0.089**	0.014
5	Silt	**0.007**	0.008	0.091	0.050	0.079	0.146
6	Sand	0.006	0.008	0.062	0.024	0.083	0.013
7	Clay	0.002	0.011	0.099	0.029	**0.085**	0.014
8	CaCo₃	**0.015**	0.010	0.089	0.039	**0.099**	0.011
9	ω	**0.008**	0.008	0.077	0.031	0.083	0.014
10	D	0.0005	0.009	**0.092**	0.037	0.079	0.011

EC: electrical conductivity; OC: organic carbon; ω: composite scaling constant; D: fractal dimension of the particle size distribution (PSD). Bold values indicate the more effective variables for prediction of N, P and K
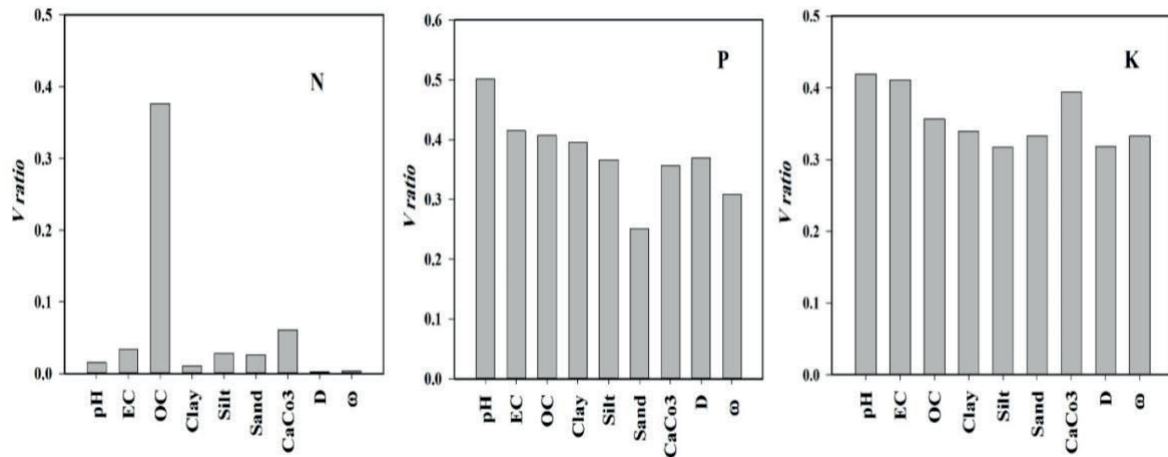
Figure 2. Sensitivity analysis of the output variables (nitrogen, phosphorus and potassium nutrients) to the input attributes of the models. EC: electrical conductivity; OC: organic carbon; D: fractal dimension of the particle size distribution (PSD); ω: composite scaling constant.

is the most important variable for the prediction of K and P, while SOM is the most important variable for the prediction of N.

Comparison of models

ANN, *k*-NN and SVR models ran using the best input attributes which were selected from the Gamma sensitivity analyses. Three-layer ANN including input layer, hidden layer, and output layer were trained for prediction of soil macronutrients. In the input layer, the number of neurons were defined as the number of independent variables (*i.e.* five variables). In the output layer, the number of neurons were determined considering the number of dependent variables (*i.e.* soil macronutrients). Amini *et al.* (2005) reported that enormous neurons can lead over fitting, whilst too few hidden neurons can cause under fitting. In this study, for finding the most equitable number of neurons in the hidden layer, neural networks with various combinations of the neuron numbers (ranging from 2 to 20) were investigated. The most suitable number of hidden neurons to predict N and K varied from 6 to 9, while the best number of hidden neurons for that of P varied between 6 and 10.

Table 3 shows the statistical summary of the studied ANN, *k*-NN and SVR models. The results indicated that the SVR was the best model in prediction of all the studied nutrients with the lowest median of RMSE. Also comparison of median of RMSE among 10 studied subsets showed that the SVR (median RMSE=3.63) and the *k*-NN (median RMSE=3.66) models had similar accuracies in the prediction of P nutrient. The *k*-NN model had smaller RMSE than the ANN for the prediction of P. But the RMSE of ANN in predicting N and K was smaller than that of the *k*-NN. Comparison of the models' performance among the studied soil nutrients displayed that the most accurate result was obtained for the N prediction models,

Table 3. Statistical summary of supervised model performances for 10 subsets

|	Median of statistical metrics	N	P	K
ANN	R	0.89	0.83	0.65
	RMSE	0.79	4.36	198
	r-RMSE	0.41	0.67	0.33
k-NN	R	0.92	0.86	0.71
	RMSE	0.93	3.66	199
	r-RMSE	0.43	0.65	0.33
SVR	R	0.95	0.83	0.79
	RMSE	0.61	3.63	170
	r-RMSE	0.29	0.64	0.29

ANN: artificial neural network; *k*-NN: *k*-nearest neighbors; SVR: support vector regression; R: correlation coefficient; RMSE: root mean square error; r-RMSE: relative root mean square error.

which might be linked to the relationship among the input and the output variables.

Measured and estimated values of the studied nutrients (N, P, and K) are compared in Fig. 3. As it can be illustrated from the scatterplots, the SVR has the nearest estimates to the measured values of all the three nutrients. Fig. 3 shows that, for N nutrient, all the applied models illustrated the least scattered estimates and their estimates were close to the measured N values compared to those of K and P nutrients. For the prediction of N, the *k*-NN method gave the worst estimates. In this case, RMSE of the SVR model was 0.63 mg g^{-1}, which was lower than that of the ANN (0.91 mg g^{-1}) and the *k*-NN (0.97 mg g^{-1}) models (Table 4). This means that the SVR model had 31% and 35% less RMSE than the ANN and the *k*-NN models, respectively. According to Table 4, the values of
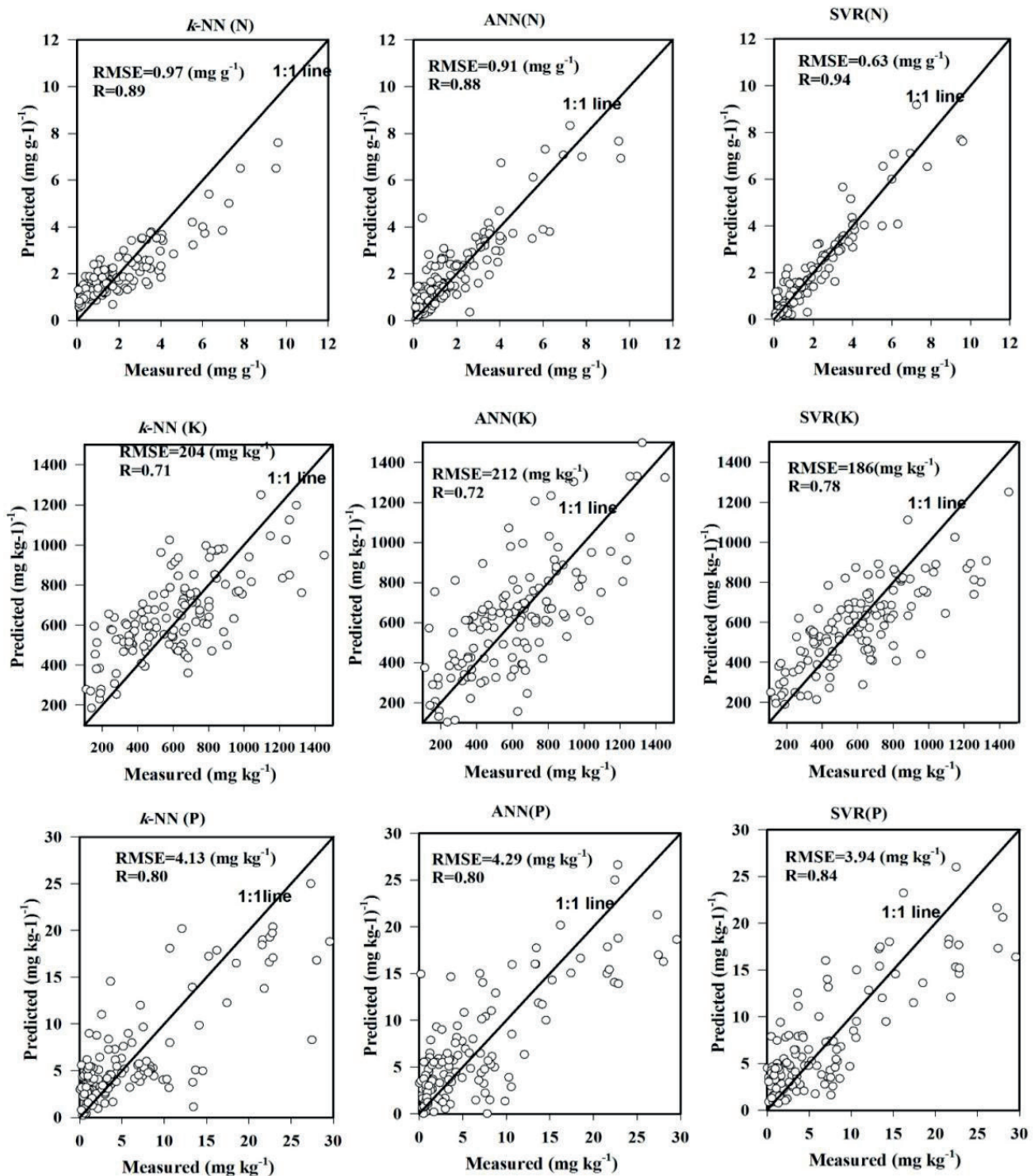
Figure 3. Scatter plots of the observed and simulated macronutrients (nitrogen, phosphorus and potassium). *k*-NN: *k*-nearest neighbors; ANN: artificial neural network; SVR: support vector regression; RMSE: root mean square error; R: correlation coefficient.

Table 4. Statistical summary of supervised model performances for the whole data set

	Model estimate	N	P	K
ANN	R	0.88	0.80	0.72
	RMSE	0.91	4.29	212
	MR	-0.13	-0.49	-13.91
k-NN	R	0.89	0.80	0.71
	RMSE	0.97	4.31	204.4
	MR	-0.002	0.062	-39.95
SVR	R	0.94	0.84	0.78
	RMSE	0.63	3.94	186
	MR	-0.008	-0.37	23.44

ANN: artificial neural network; *k*-NN: *k*-nearest neighbors; SVR: support vector regression; R: correlation coefficient; RMSE: root mean square error; MR: mean error.

RMSE for ANN, *k*-NN, and SVR for K were 212, 204.4 and 186 (mg kg^{-1}), respectively. These results demonstrate that the *k*-NN model had lower RMSE compared to the ANN model. For K, the SVR model had 12% and 8% less RMSE than ANN and *k*-NN models, respectively. For P, SVR also revealed a lower RMSE compared to *k*-NN and ANN models but both ANN and *k*-NN models had similar RMSE. In this case, the SVR model had 8% less RMSE than both ANN and *k*-NN models. MR for estimating N and P values using the three models was negative (Table 4). These findings indicate that the studied models overestimated the values of P and N. For K, the ANN and *k*-NN models had negative MR but SVR had a positive MR, showing that both ANN and *k*-NN models overestimate the K values but SVR underestimates the K values. Zolfaghari *et al.* (2016) reported that the bias in *k*-NN in the prediction of CEC was positive. Nemes *et al.* (2006) also reported a positive MR for estimation of soil moisture content using *k*-NN.

Fig. 4 illustrates r-RMSE values split up per test stage for all three nutrients and models. For N, the individual accuracy of the ANN model per test stages fluctuated more than *k*-NN and SVR models. But the *k*-NN model approach illustrated the maximum r-RMSE (0.76) value (minimum degree of accuracy). For K, the values from the ANN models fluctuated more than *k*-NN and SVR models. For this nutrient, the maximum r-RMSE (0.46) was obtained by both *k*-NN and ANN models. For P, r-RMSE fluctuation of all the studied models was the same. Our results indicate that the SVR tends to be a more accurate model for prediction of macronutrients. These results were different from Shiri *et al.* (2017), who reported that support vector machine (SVM) gives the worst estimates of soil CEC. The performance fluctuations among test stages indicate the need for evaluation of the models' performances using *k*-fold testing data set assignment approach and not only considering a single data set assignment.
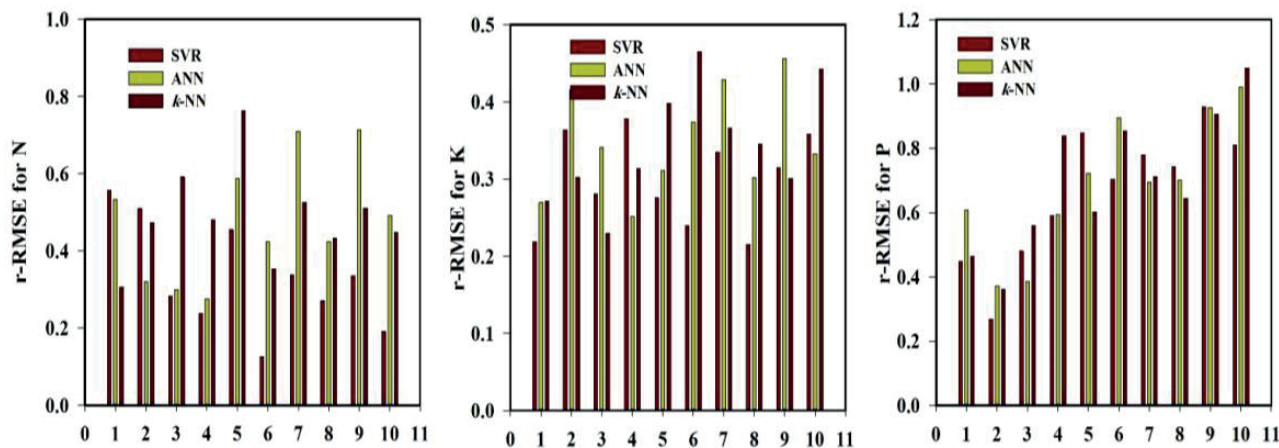
Figure 4. Relative root mean square error (r-RMSE) values for the three soil nutrients (nitrogen, phosphorus and potassium) estimates split up per test stage. SVR: support vector regression; ANN: artificial neural network; *k*-NN: *k*-nearest neighbors.

In summary, easily measurable soil properties, including soil PSD, pH, EC, CaCO$_3$, OC and fractal parameters (D and ω) were used to predict soil macro-nutrients (*i.e.* N, P, and K) using nonparametric methods like ANN, *k*-NN and SVR. Sensitivity analysis obtained by Gamma test showed that OC, CaCO$_3$, EC, ω, and silt for N; pH, EC, OC, clay and D for P; pH, EC, CaCO$_3$, OC and clay for K increased the Gamma value and subsequently had profound positive effects on the performance of the studied models. Results demonstrated that SVR has the best performance for the prediction of all studied nutrients. Also, there was no significant difference in accuracy of SVR and *k*-NN in prediction of P. In *k*-NN method the user can incorporate additional data by appending to or replacing the reference database without the need for developing new equations. In addition, the user is able to improve a specific local data by inserting available local data to a reference database without any significant influence on other available parts of the data, in the reference database. Therefore, we suggest to use the *k*-NN approach for prediction of P nutrient. There was no significant difference in the performance of *k*-NN and ANN models in prediction of N and K nutrients. Thus, we conclude that the *k*-NN approach could well compete with many other methods of pedotransfer functions, like ANN models. The lowest and the highest uncertainties were found in the prediction of N and K nutrients, respectively.

Acknowledgments

References

Adviento-Borbe M, Doran J, Drijber R, Dobermann A, 2006. Soil electrical conductivity and water content affect nitrous oxide and carbon dioxide emissions in intensively managed soils. J Environ Qual 35: 1999-2010. https://doi.org/10.2134/jeq2006.0109

Aimrun W, Amin M S, Rusnam M, Ahmad D, Hanafi M, Anuar A, 2009. Bulk soil electrical conductivity as an estimator of nutrients in the maize cultivated land. Eur J Sci Res 31: 37-51.

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

Ayoubi S, Sahrawat KL, 2011. Comparing multivariate regression and artificial neural network to predict barley production from soil characteristics in northern Iran. Arch Agron Soil Sci 57: 549-565. https://doi.org/10.1080/03650341003631400

Ballabio C, 2009. Spatial prediction of soil properties in temperate mountain regions using support vector regression. Geoderma 151: 338-350. https://doi.org/10.1016/j.geoderma.2009.04.022

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

Babaei F, Zolfaghari A A, Yazdani M R, Sadeghipour A, 2018. Spatial analysis of infiltration in agricultural lands in arid areas of Iran. Catena 170: 25-35. https://doi.org/10.1016/j.catena.2018.05.039

Bird N, Perrier E, Rieu M, 2000. The water retention function for model of soil structure with pore and solid fractal distribution. Eur J Soil Sci 51: 55-56. https://doi.org/10.1046/j.1365-2389.2000.00278.x

Botula Y, Nemes A, Mafuka P, Ranst E, Cornelis W, 2013. Prediction of water retention of soils from the humid tropics by the nonparametric *k*-nearest neighbor approach. Vadose Zone J 12: 1-17. https://doi.org/10.2136/vzj2012.0123

Bronson KF, Booker J, Keeling JW, Boman RK, Wheeler TA, Lascano RJ, Nichols RL, 2005. Cotton canopy reflectance at landscape scale as affected by nitrogen fertilization. Agron J 97: 654-660. https://doi.org/10.2134/agronj2004.0093

Chaudhari P, Ahire D, Ahire VD, 2012. Correlation between physico-chemical properties and available nutrients in sandy loam soils of haridwar. J Chem Biol Phys Sci 2: 1493.

Cherkassky V, Mulier FM, 2007. Learning from data: concepts, theory, and methods. John Wiley & Sons, NJ, USA. https://doi.org/10.1002/9780470140529

Foth H, 1982. Soil resources and food: A global view. In: Principles and applications of soil geography; Bridges EM & Davidson DA (eds), pp: 48-61. Longman, NY.

Fu W, Tunney H, Zhang C, 2010. Spatial variation of soil nutrients in a dairy farm and its implications for site-specific fertilizer application. Soil Till Res 106: 185-193. https://doi.org/10.1016/j.still.2009.12.001

Gee G, Bauder J, 1986. Particle size analysis. In: Methods of soil analysis, part 1; Klute A (ed), pp: 383-411. Agron. Monogr. No. 9. Am Soc Agron, Madison, WI, USA. https://doi.org/10.2136/sssabookser5.1.2ed.c15

Haykin S, 1994. Neural networks: a comprehensive foundation. Prentice Hall PTR.

Jackson ML, 2005. Soil chemical analysis: Advanced course. UW-Madison Libraries Parallel Press.

Jeong G, Oeverdieck H, Park SJ, Huwe B, Ließ M, 2017. Spatial soil nutrients prediction using three supervised

learning methods for assessment of land potentials in complex terrain. Catena 154: 73-84. https://doi.org/10.1016/j.catena.2017.02.006

Keshavarzi A, Sarmadian F, Omran ESE, Iqbal M, 2015. A neural network model for estimating soil phosphorus using terrain analysis. Egypt J Remote Sens Space Sci 18: 127-135. https://doi.org/10.1016/j.ejrs.2015.06.004

Kovačević M, Bajat B, Gajić B, 2010. Soil type classification and estimation of soil properties using support vector machines. Geoderma 154: 340-347. https://doi.org/10.1016/j.geoderma.2009.11.005

Kuhn M, Johnson K, 2013. Applied predictive modeling. Springer. https://doi.org/10.1007/978-1-4614-6849-3

Li QQ, Yue TX, Wang CQ, Zhang WJ, Yu Y, Li B, Yang J, Bai GC, 2013. Spatially distributed modeling of soil organic matter across china: An application of artificial neural network approach. Catena 104: 210-218. https://doi.org/10.1016/j.catena.2012.11.012

Li QQ, Zhang X, Wang CQ, Li B, Gao XS, Yuan DG, Luo YL, 2016. Spatial prediction of soil nutrient in a hilly area using artificial neural network model combined with kriging. Arch Agron Soil Sci 62: 1541-1553. https://doi.org/10.1080/03650340.2016.1154543

Marti E, Jofre J, Balcazar JL, 2013. Prevalence of antibiotic resistance genes and bacterial community composition in a river influenced by a wastewater treatment plant. PLoS One 8 (10): e78906. https://doi.org/10.1371/journal.pone.0078906

Mathworks, 2010. Matlab Version 7.0. The Mathworks Inc., Natick, MA, USA.

Moghaddamnia A, Gousheh MG, Piri J, Amin S, Han D, 2009. Evaporation estimation using artificial neural networks and adaptive neuro-fuzzy inference system techniques. Adv Water Resour 32: 88-97. https://doi.org/10.1016/j.advwatres.2008.10.005

Moosavi AA, Sepaskhah A, 2012. Artificial neural networks for predicting unsaturated soil hydraulic characteristics at different applied tensions. Arch Agron Soil Sci 58: 125-153. https://doi.org/10.1080/03650340.2010.512289

Nemes A, Rawls WJ, Pachepsky YA, 2006. Use of the nonparametric nearest neighbor approach to estimate soil hydraulic properties. Soil Sci Soc Am J 70: 327-336. https://doi.org/10.2136/sssaj2005.0128

Olsen SR, Khasawneh FE, 1980. Use and limitations of physical-chemical criteria for assessing the status of phosphorus in soils. In: The role of phosphorus in agriculture; Khasawneh FE, Sample EC & Kamprath EC (eds), pp: 361-410. ASA, CSSA, and SSSA Books. https://doi.org/10.2134/1980.roleofphosphorus.c15

Peralta NR, Costa JL, 2013. Delineation of management zones with soil apparent electrical conductivity to improve nutrient management. Comput Electron Agric 99: 218-226. https://doi.org/10.1016/j.compag.2013.09.014

Raheb A, Heidari A, 2012. Effects of clay mineralogy and physico-chemical properties on potassium availability under soil aquic conditions. J Soil Sci Plant Nutr 12: 747-761. https://doi.org/10.4067/S0718-95162012005000029

Safaa M, Maxwellb TMR, 2015. Predicting pasture nitrogen content using ANN models and thermal images. 21st Int Congr on Modelling and Simulation, Gold Coast, Australia.

Shiri J, Nazemi AH, Sadraddini AA, Landeras G, Kisi O, Fard AF, Marti P, 2014. Comparison of heuristic and empirical approaches for estimating reference evapotranspiration from limited inputs in Iran. Comput Electron Agric 108: 230-241. https://doi.org/10.1016/j.compag.2014.08.007

Shiri J, Keshavarzi A, Kisi O, Iturraran-Viveros U, Bagherzadeh A, Mousavi R, Karimi S, 2017. Modeling soil cation exchange capacity using soil parameters: Assessing the heuristic models. Comput Electron Agric 135: 242-251. https://doi.org/10.1016/j.compag.2017.02.016

Simon HA, 1995. Artificial intelligence: An empirical science. Artif Intell 77: 95-127. https://doi.org/10.1016/0004-3702(95)00039-H

Skaggs T, Arya L, Shouse P, Mohanty B, 2001. Estimating particle-size distribution from limited soil texture data. Soil Sci Soc Am J 65: 1038-1044. https://doi.org/10.2136/sssaj2001.6541038x

Smola AJ, Schölkopf B, 2004. A tutorial on support vector regression. Stat Comp 14 (3): 199-222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

Taghizadeh-Mehrjardi R, Toomanian N, Khavaninzadeh A, Jafari A, Triantafilis J, 2016. Predicting and mapping of soil particle-size fractions with adaptive neuro-fuzzy inference and ant colony optimization in central Iran. Eur J Soil Sci 67: 707-725. https://doi.org/10.1111/ejss.12382

Vapnik V, 1995. The nature of statistical learning theory. Springer, NY. https://doi.org/10.1007/978-1-4757-2440-0

Walkley A, Black IA, 1934. An examination of the degtjareff method for determining soil organic matter, and a proposed modification of the chromic acid titration method. Soil Sci 37: 29-38. https://doi.org/10.1097/00010694-193401000-00003

Wang H, Shi X, Yu D, Weindorf DC, Huang B, Sun W, Ritsema CJ, Milne E, 2009. Factors determining soil nutrient distribution in a small-scaled watershed in the purple soil region of Sichuan province, China. Soil Till Res 105: 300-306. https://doi.org/10.1016/j.still.2008.08.010

Were K, Bui DT, Dick ØB, Singh BR, 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. https://doi.org/10.1016/j.ecolind.2014.12.028

Yazdani MR, Zolfaghari AA, 2017. Monthly river forecasting using instance-based learning methods and climatic parameters. J Hydrol Eng 22 (6): 04017002. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001490

Zolfaghari A, Taghizadeh-Mehrjardi R, Moshki A, Malone B, Weldeyohannes A, Sarmadian F, Yazdani M, 2016. Using the nonparametric *k*-nearest neighbor approach for predicting cation exchange capacity. Geoderma 265: 111-119. https://doi.org/10.1016/j.geoderma.2015.11.012