Geoderma 265 (2016) 111–119

Contents lists available at ScienceDirect

Geoderma

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

Using the nonparametric *k*-nearest neighbor approach for predicting cation exchange capacity

A.A. Zolfaghari [a,*], R. Taghizadeh-Mehrjardi [b], A.R. Moshki [a], B.P. Malone [c], A.O. Weldeyohannes [d], F. Sarmadian [e], M.R. Yazdani [a]

[a] Faculty of Desert studies, Semnan University, Semnan, Iran
[b] Faculty of Agriculture and Natural Resources, Ardakan University, Ardakan, Iran
[c] Soil Security Laboratory, Faculty of Agriculture & Environment, The University of Sydney, NSW 2006, Australia
[d] Department of Renewable Resource, University of Alberta, Alberta, Canada
[e] Department of soil science, University of Tehran, Karaj, Iran

ARTICLE INFO

Article history:
Received 25 February 2015
Received in revised form 15 July 2015
Accepted 8 November 2015
Available online 1 December 2015

Keywords:
Nonparametric k-nearest neighbor (*k*-NN)
Artificial neural network
Cation exchange capacity (CEC)
Iran

ABSTRACT

The objectives of this study were to apply a *k*-NN approach to predict CEC in Iranian soils and compare this approach with the popular artificial neural network model (ANN). In this study, a data set of 3420 soil samples from different parts of Iran was used. Two different sets of cheaper-to-measure soil attributes were selected as potential predictors. The first set consisted of clay, silt, sand and organic carbon (OC) contents. The second data set was constructed using OC and clay contents. Two 'design-parameter' parameters should be optimized before application of the *k*-NN approach. Results showed that the algorithm efficiency is not dependent on these parameters. A wide range of suboptimal values around the optimal values may cause a slight error in terms of estimation accuracy. However, the optimal settings of the design-parameters depend on the size of the development/reference data set. In both *k*-NN and ANN models, the higher number of input variables can relatively improve the estimation of CEC. But this improvement was not statistically significant at the 0.05 level. Furthermore, the results showed that increasing the size of the reference data set to a certain amount ($N = 1200$) reduced the estimation error significantly in terms of root-mean-squared residuals (RMSE). However, no significant difference in the accuracy of *k*-NN and ANN methods was detected in the reference data set sizes for $N > 1200$. Results showed no significant difference between this approach and ANN models, suggesting the competitive advantage of the *k*-NN technique over other techniques to develop pedotransfer functions (PTFs), for example, the redevelopment of PTFs is not necessarily required as new data become available.

1. Introduction

Cation exchange capacity (CEC) is one of the most important soil indicators which controls some basic attributes such as soil acidity, nutrient retention capacity and environmental quality. Therefore, it is considered as one of the key parameters of soil fertility and productivity management (Krogh et al., 2000). As a result, CEC has long been an input parameter of many environmental models (Keller et al., 2001). Added to this, CEC data can give clear and complete interpretation of soil and plant nutrition processes, and consequently fertilizer and soil amendment requirements. Laboratory analysis is the most accurate method for direct measurement of CEC. However, direct measurement of CEC is difficult, particularly in arid and semi arid region soils of Iran, due to large amounts of calcium carbonate that makes measurement expensive and time-consuming (Seybold et al., 2005; McBratney et al., 2002; Amini et al., 2005).

Pedotransfer functions (PTFs) are indirect methods that have been used to estimate hard-to-measure soil properties from easier measured and often readily available soil properties. In recent decades, several PTFs have been used to developed, correlating CEC to more readily available soil data such as soil texture, organic carbon and pH. The CEC of soils is usually related to soil texture, organic matter, mineralogy and the fractal dimensions of particle size distribution (Bayat et al., 2014). Krogh et al. (2000) used PTFs to predict CEC of soils and found that 90% of the variation of CEC was attributed to soil clay and organic matter content. Similar results were obtained by Bell and van Keulen (1995), who showed that more than 96% of CEC variation in the soil could be explained by clay, organic matter contents and pH values.

Regarding regression PTFs as a parametric approach, each parameter of the regression should be calibrated by an optimal fitting of that equation to data. Moreover, new parameters must be adjusted for each series of new data and assumptions are required for variable distribution. However, the model may be difficult to validate if the data set is small. Small data sets could also lead to issues of model bias (Nemes et al., 2006). To overcome these limitations, recent research findings have

* Corresponding author.
E-mail address: azolfaghari@semnan.ac.ir (A.A. Zolfaghari).

http://dx.doi.org/10.1016/j.geoderma.2015.11.012
0016-7061/

suggested non-parametric methods such as artificial neural networks (ANNs) as ideal. Applications of ANNs to soil science are varied ranging from determining soil moisture (Frate et al., 2003), field capacity and permanent wilting point (Nemes et al., 2006), to the development of PTFs for prediction of CEC. Amini et al. (2005) indicated that the ANN method can improve accuracy of CEC prediction up to 25% in comparison with multiple linear or least-squares regression methods. Another form of non-parametric method is *k*-nearest neighbor (*k*-NN), a technique substantially based on the principles of similarity and proximity of data. The *k*-NN method is widely used in agriculture (Bannayan and Hoogenboom, 2009), forestry (Lopez et al., 2001) and hydrology (Clark et al., 2004; Yates et al., 2003). It is one of such approaches which have been applied to estimate soil physical and chemical characteristics and would be more useful when the relationship between input and output data is not clear (Nemes et al., 2006). Jalali and Homaee (2011) applied this approach for estimating saturated soil hydraulic conductivity using particle size distribution, bulk density, organic carbon, electrical conductivity, and saturation soil moisture content. They reported that this technique has a good ability to estimate a given target variable, and it can be considered as a good candidate model for PTFs. Nemes et al. (2006) estimated the soil moisture content with matric potentials −33, −1500 kPa soil suction, from particle size distribution, bulk density and soil organic matter data using ANN and *k*-NN approaches. The two approaches have the same accuracy for estimation and derivation transfer functions (Nemes et al., 2006). Botula et al. (2013) also used *k*-NN approach to predict soil water retention in a humid tropical region and showed that the estimation error in *k*-NN approach is lower than the examined multiple linear regression PTFs. Soil hydraulic properties can better be estimated by the non-parametric method *k*-NN compared to the parametric PTFs (Nemes et al., 2008). Furthermore, Haghverdi et al. (2010) showed that the *k*-NN method can more efficiently estimate the soil moisture in matric potentials −33 and −1500 kPa compared to the ANN method in the north and northeast soils of Iran. However, it seems that the application of *k*-NN for prediction of CEC is quite new and an investigation is warranted.

Therefore, the objectives of this study were (i) to apply a non-parametric approach to estimate CEC using an adaptation of the *k*-NN algorithm developed by Nemes et al. (2006); (ii) to test the ability of the *k*-NN algorithm to predict CEC of arid and semiarid soils of Iran; and (iii) to compare the prediction performance of the proposed *k*-NN variant with the more common ANN model.

2. Material and methods

2.1. Data review

In Iran, a total of 15 World Resources Base (WRB) groups have been recorded. These include: Calcisols, Cambisols, Chernozems, Fluvisols, Gleysols, Gypsisols, Kastanozems, Leptosols, Luvisols, Phaeozems, Regosols, Solonchaks, Solonetz, and Vertisols. The most frequently observed soil classes are: Regosols, Calcisols, and Gypsisols. Most parts of the central plains of Iran are not suitable for cultivation except in oasis area, around the cities and in mountainous plains. The northern Caspian coast of Iran is the most cultivated part of the country (Hengl et al., 2007). Illite, high charge smectite, palygorskite, chlorite, vermiculite and kaolinite silicate clay minerals were found in almost all the soils studied. However, illite and chlorite clay minerals were the dominant in arid regions of Iran. The origin of illite and chlorite was mainly from the parent materials. The amount of vermiculite and smectite were higher in the soils of northern Iran developed under humid condition. Palygorskite was found in the higher amounts in the lower areas under the saline and sodic condition. Palygorskite was a dominant clay mineral in gypsiferous and calcareous Aridisols. In this paper a collection of 3420 soil samples from around Iran representing different parts of Iran were used as the reference or training data set of *k*-NN and ANN estimations (Fig. 1.).

The soil samples were dried, crushed and passed through a 2 mm sieve to prepare for physical and chemical analysis. The percentages of sand (50–2000 μm), silt (2–50 μm) and clay (<2 μm) were determined using the hydrometer method (Gee and Bauder, 1986) according to the USDA soil textural classification system. Fig. 2 shows the textural distribution of the data set. The soil organic carbon was determined using the Walkly–Black method (Nelson and Sommers, 1982) and the CEC was determined by the ammonium saturation method at pH 7.0 (Soil Survey Staff, 1993).

2.2. Reference and test data set

The data set was randomly partitioned into calibration and test sets. Here 720 samples were designated as test data with the remaining 2700 used for PTF calibration. For a sensitivity analysis calibration data sets of size: 100, 200, 400, 800,1200, 1600, 200 and 2500 samples were also used in order to compare whether sample size is important to the development of PTFs using either the *k*-NN and ANN modeling approaches. Further to this, all the random data selections were repeated 50 times to allow the development of an ensemble of PTF estimations and subsequent estimation of prediction uncertainty.

In this study, the following two data sets were used as model inputs for predicting CEC. The first set consisted of clay, silt, sand and organic carbon (OC) contents. The second data set was constructed using OC and clay contents.

2.3. Prediction models

2.3.1. The k-nearest neighbor technique (k-NN)

The *k*-NN algorithm used in this study was adapted from the variant developed by Nemes et al. (2006) and it was implemented in the MATLAB environment (Mathworks, 2010). The *k*-NN technique does not use any predefined mathematical functions to estimate a target variable. A reference data set is searched for soils that are most similar to the target soil on the basis of the selected input attributes or features. Apparently, the performance of technique largely depends on the goodness of selection of the 'most similar' (nearest) soils. The similarity between the target soils and the known instances was measured in terms of a metric considered here as the Euclidean distance:

$$d_i = \sqrt{\sum_{j=1}^{x} \Delta a_{ij}{}^2} \tag{1}$$

where: d_i is the "distance" of the *i*th soil from the target soil and Δa_{ij} is the difference of the *i*th soil from the target soil in the *j*th soil attribute and *x* is the number of soil properties considered for the model. The term 'distance', does not refer to actual (physical) distance, but to a measure of similarity; the distance will be smaller for soils that are more similar to the target soil in regard to the input attributes.

Soils show some attributes that differ in orders of magnitude or range. For example, sand content varied between 1% to 98%, but OC varied between 0.009% to 8.8%. Therefore, a unit difference in OC is expected to be more significant than the same unit difference in sand content. Therefore, a normalization procedure was applied to the soil attributes data before they were used to calculate the Euclidean distance in Eq. (1). Firstly, all the input attributes were first transformed to temporary variables $a_{ij(temp)}$ with a distribution having zero mean and one standard deviation as follows:

$$a_{ij(temp)} = \frac{\left[(a_{ij}) - \overline{a}_i\right]}{\sigma(a_j)} \tag{2}$$

where a_{ij} is the value of the jth attribute of the *i*th soil, and $\bar{a}_j$ and $\sigma(a_j)$ are the mean and standard deviation of the observed values of the *j*th attribute in the reference data set. Secondly, the difference

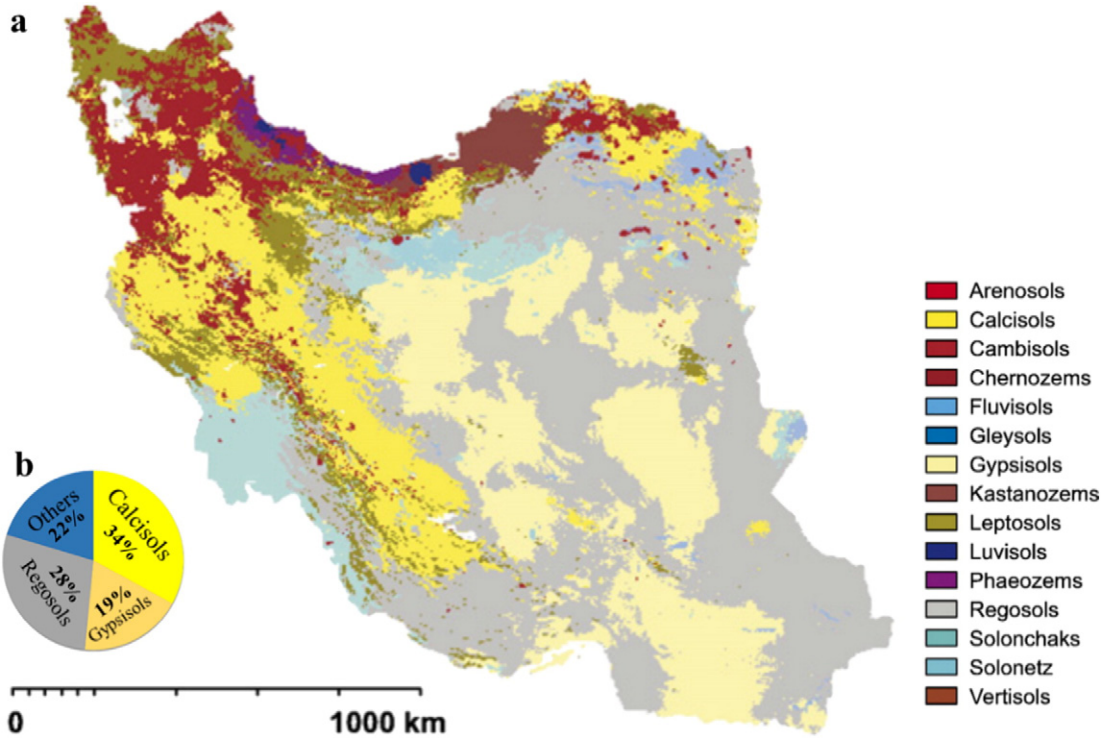

Fig. 1. Spatial distribution of WRB soil groups (a) and percentage of soil samples collected in the most frequent classes (b). After Hengl et al., 2007.

between the minimum and maximum of these temporary variables were then examined to identify the soil attribute that showed the widest range of transformed (temporary) values. This allowed for obtaining zero mean and the same minimum–maximum range in the data of all attributes:

$$a_{ij} = \frac{\{\max[range(a_{j=1(temp)}), \ldots range(a_{j=x(temp)})]\}}{range(a_{j(temp)})} \quad (3)$$

where $a_j(temp)$ is the data of the jth soil attribute normalized using Eq. (2) and $a_{ij}(trans)$ is the final transformed value of the jth attribute of the ith soil. Eventually, a_{ij} $(trans)$ values derived from Eq. (3) were used as input in k-NN algorithm.

Finally, one has to decide how to weigh each selected soil while forming the estimate of the output attribute. The reference data set soils were sorted in ascending order of their distance to the target soil. A weighting procedure that accounts for the distribution of the distances of the selected k-neighbors from the target soil was applied. Weights of each selected neighbor were computed as:

$$w_i = \frac{d_{i(rel)}}{\sum_{i=1}^{k} d_{i(rel)}} \quad (4)$$

where: k is the number of neighbors selected, w_i is the weight associated with the ith nearest neighbor, and $d_i(rel)$ is the relative distance of the ith selected neighbor, calculated as:

$$d_{i(rel)} = \left(\frac{\sum_{i=1}^{k} d_i}{d_i} \right)^{p} \quad (5)$$

where: d_i is the distance of the ith selected neighbor computed using Eq. (1) and p is a power term to account for different possible

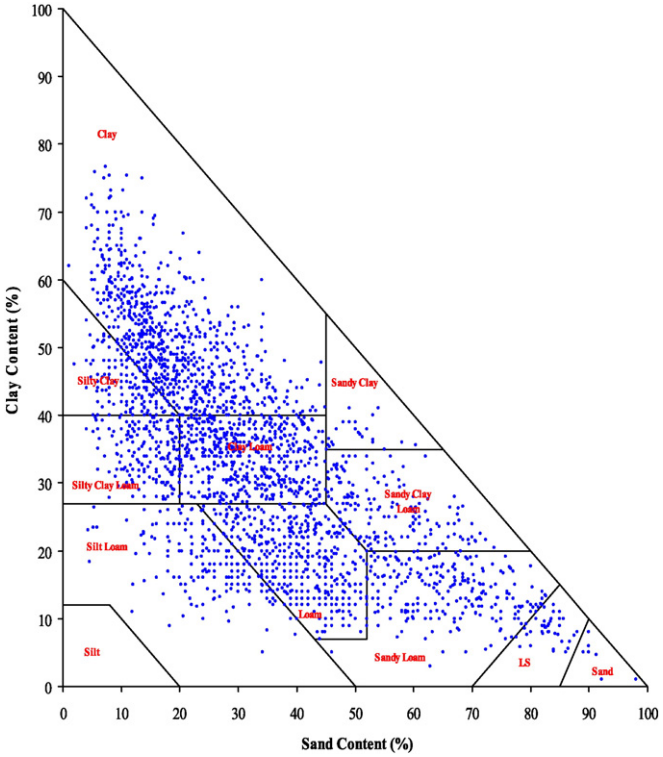

Fig. 2. Variation of clay, silt, and sand in the data set.

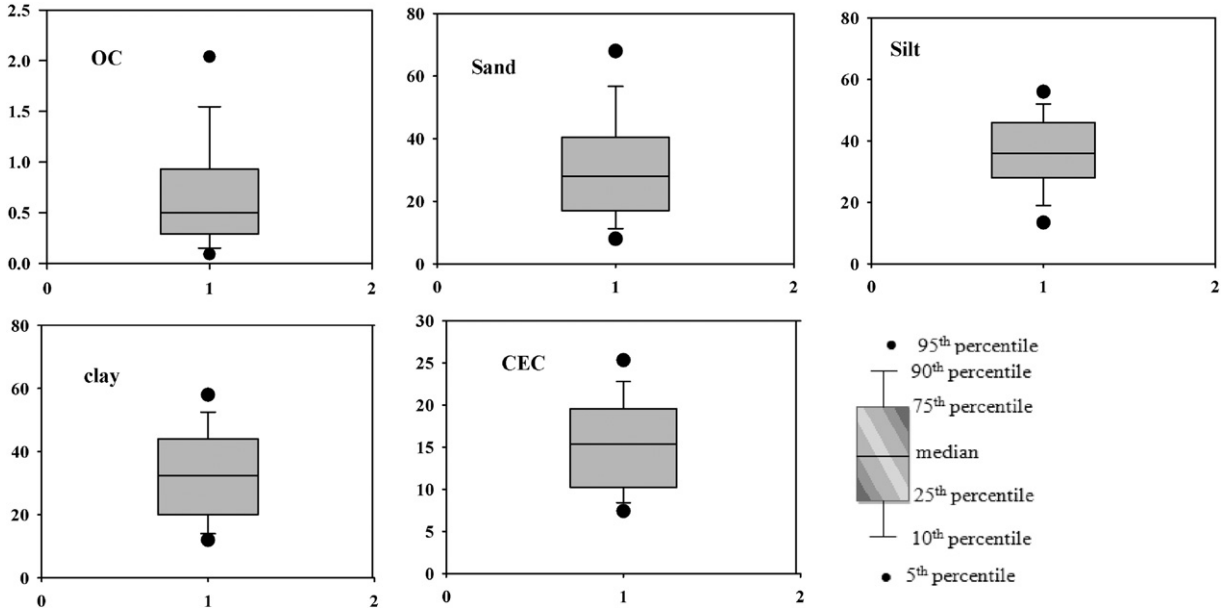

Fig. 3. Box plots of organic content (OC, %), sand (%), silt (%), clay (%) and cation exchange capacity (CEC, cmol kg^{-1} soil) of the data set.

weight–distance relationships. Therefore, CEC was predicted as:

$$\hat{CEC}_i = \sum_{i=1}^{k} w_i CEC_i \tag{6}$$

where: $\hat{CEC}_i$ is predicted CEC of *i*th soil in test data set and CEC_i is *i*th soil sample in reference data set.

2.3.2. Artificial neural network model (ANN)

The ANN model with three-layer back propagation ANN model was used to estimate the target data. Various methods were presented to determine the number of neurons in the hidden layer. For example, the integer rounded up half the total number of input and output variables can be used as the number of neurons in the hidden layer (Nemes et al., 2006). Subsequently, this protocol was adopted for this study. Before running the program, all the data were scaled to occur between 0 and 1 in order to increase accuracy of the results. The ANN modeling was performed using the Neural Network Toolbox in MATLAB (Mathworks, 2010).

2.4. Sensitivity analysis

Sensitivity analysis was performed using regression methods to evaluate the effect of each of the input variables on the estimated soil CEC using *k*-NN and ANN models. The principle of regression methods is to approximate mapping between an output and the factors by an equation of the form:

$$y = b_0 \sum_{i=1}^{n} b_i x_i + \varepsilon \tag{7}$$

where: y is the model output, x_i is the *i*th model input, *n* is the number of input, b_i is the coefficient to be estimated for each x_i, and ε is random error.

When the input x_i are independent of one another, then the standardized regression coefficient (SRC) can be used to provide a sensitivity index for the input x_i:

$$SRC(i) = b_i \frac{\widehat{s}_i}{\bar{s}} \tag{8}$$

where: $\widehat{s}_i$ is the input standard deviation and *s* the output standard deviation.

Each of the SRC gives information about the effect of changing the value of an input from its standard value by a fixed fraction of its standard deviation, while maintaining the other factors at their default values. If the regression is actually able to explain the data, the larger the SRC value the more sensitive the model output is to the input variables (Confalonieri et al., 2010).

2.5. Evaluation of models:

To evaluate the efficiency and accuracy of *k*-NN and ANN models, the root mean square error (RMSE) and mean error (MR) criteria were used. The mean error indicates whether there is a systematic error in the method and the RMSE shows the accuracy of the method. Both

Table 1
Statistics of the data set.

Properties	Unit	Min	Max	Ave	SD	Median	CV	Skewness
USDA sand	%	1.00	98.00	31.01	17.83	28.00	0.58	0.93
USDA silt	%	1.00	77.20	36.10	12.78	36.00	0.36	−0.14
USDA clay	%	1.00	76.70	32.90	14.75	32.50	0.45	0.28
OC	%	0.009	8.80	0.72	0.70	0.50	0.97	3.31
CEC	$cmol_c\ kg^{-1}$ soil	1.85	39.60	15.38	5.78	15.36	0.38	0.31

Table 2
Correlation coefficients (r) of the measured soil attributes.

	Sand	Silt	Clay	OC	CEC
Sand	1**	−0.57**	−0.70**	−0.11**	−0.48**
Silt		1**	−0.17**	0.12**	−0.09**
Clay			1**	0.10**	0.70**
OC				1**	0.35**
CEC					1**

** means significant differences ($p < 0.01$)

RMSE and MR criteria were determined as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(\hat{CEC}_i - CEC_i\right)^2}{N}} \quad (9)$$

$$MR = 1/N\sum_{i=1}^{N}\left(CEC_i - \hat{CEC}_i\right) \quad (10)$$

where: CEC_i, $\hat{CEC}_i$are observed and predicted CEC for ith soil sample, respectively and N is the number of samples in the test data set.

3. Results

3.1. Summary statistics

Box plots of selected soil attributes for the data set are given in Fig. 3.

The selected data set showed a wide range of soil particle size distribution (Figs. 2, 3 and Table 1). The clay fraction of soil varied between 1%–70.76%. The amount of OC ranged between 0.009%–8.8% with the average of 0.72%. Soil OC content, as reflected by the coefficient of variation, showed the highest variability compared to those of the soil particle-size distribution and CEC (Table 1). The lowest coefficient of variation (CV) was for silt content. Recalling the soil samples were collected from different regions of Iran, those samples from the north parts of Iran with humid climate had higher OC. In contrast, the soils from the central parts of the country, which is characterized by a dry climate, showed low OC.

The correlation coefficients between soil attribute are given in Table 2. A strong correlation was observed between soils CEC and clay content ($r = 0.70^{**}$) (Table 2). The correlation between CEC and OC ($r = 0.35$) found in this paper was lower than those reported by Amini et al. (2005) for the arid soils of Iran ($r = 0.65$ for CEC and OC content). However, our findings agreed with those by Bayat et al. (2014) who found a weak correlation between CEC and OC ($r = 0.43$) in humid area of Iran. Considering the correlation between clay content and CEC, we found greater correlation than those values reported by Manrique et al. (1991) for the arid soils of USA ($r = 0.55$). Furthermore, CEC was negatively correlated with sand and silt contents (Table 2) which is in agreement with the findings reported by Amini et al. (2005) and Bayat et al. (2014), respectively.

3.2. Prediction models:

3.2.1. Optimizing the k and p terms

For the k-NN technique, two parameters need to be optimized. The first was the number of soil samples to be used for formulating the target soil estimation. The second parameter was the p term introduced in Eq. (5) and it was used to weigh each of the selected k soils while forming the estimate of the output attribute. The optimal parameter settings were determined by changing iteratively both of the parameters in the algorithm and making estimations of the test data set from the reference data set. For the parameter k, values from 1 to 50 by 1 increment were considered, while the parameter p was varied between 0 and 4, with 0.1 increment. Then the RMSE values were determined for each value of p and k. To avoid errors when determining the optimal values of p and k, the program was applied at different levels of p and k, and finally an average value of RMSE was used. Fig. 4. shows the average RMSE obtained from 2500 samples in the data set. It was likely that the k-NN technique showed little sensitivity to the choice of parameter p. The different values of p actually have no significant effect on the errors. For example, different values of p in optimal k vary the RMSE by about 0.05 $cmol^+ kg^{-1}$. The k-NN approach was not also very sensitive to the choice of k as long as k is above a certain minimum, which was 9 or 10. The low sensitivity of k-NN method to p and k values observed in this paper was consistent with the reports by Lall and Sharma (1996); Jagtap et al. (2004) and Nemes et al. (2006).

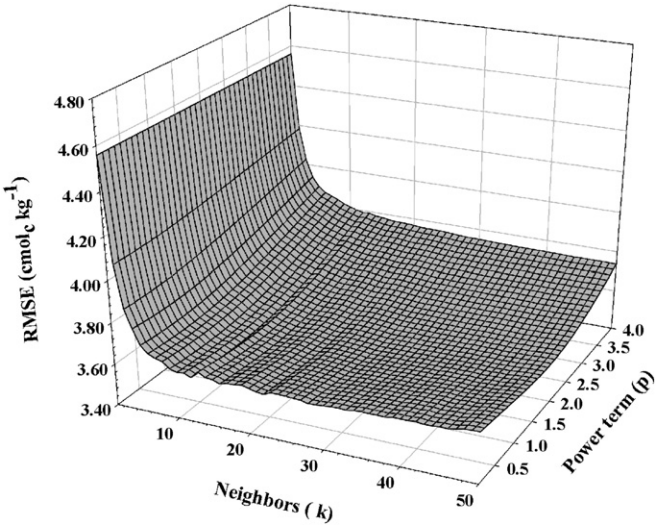

Fig. 4. Three-dimensional representation of the relationship between the number of selected neighbors, the p term used in weighing the selected neighbors and the obtained average root mean square error, using 2500 soils for each replicate of estimations.

The interdependence of the k and p terms and the number of samples in the reference data set is given in Fig. 5. Estimations developed from small data subsets (e.g., here $N = 100$ or 200) were more sensitive to changes in k and p compared to those developed from large dataset ($N > 200$). For example, including more samples from the reference data set in each individual estimation such as by increasing k beyond a threshold generally yielded the worst estimations. This was because with the small N, increasing k meant that a relatively large proportion of the data set is included in the estimation, rather than a small, but a more specific set of samples with very similar characteristics to the target sample. Hence, the estimates tend to come closer and closer to the reference data set mean, yielding less accurate "local" estimates. When k is relatively large and p is kept small, even less similar samples will have a relatively large weight in the formulation of the final CEC estimate. On the contrary, the effect of a relatively large p value, including from large sample size, on the individual estimation such as k is increased the nearest samples by their properties would receive a very high proportion of the weights while formulating the final estimate. In essence, a large p value can likely counteract the negative effect of choosing a k value that is too large. This effect was best seen when k was disproportionately high compared with N. This combined effect was less and less expressed with an increasing size of the reference data set.

Due to the low sensitivity of k-NN to the parameter p, choosing a value of 1 for p did not cause significant error in the estimation of CEC. Hence, by considering $p = 1$, Eqs. (4) and (5) can be integrated in the following simple equation, which could be used for weighting the k of a reference data set.

$$w_i = \frac{1/d_i}{\sum_{i=1}^{k} 1/d_i}. \quad (11)$$

Despite the low sensitivity of k-NN method to the values of p and k, it was necessary to determine the optimal values of the above parameters for running this algorithm. The best choice might be to use the k and p values with the minimum values of RMSE. On the other hand, the reference database can possibly affect the optimal values of k and p. For this

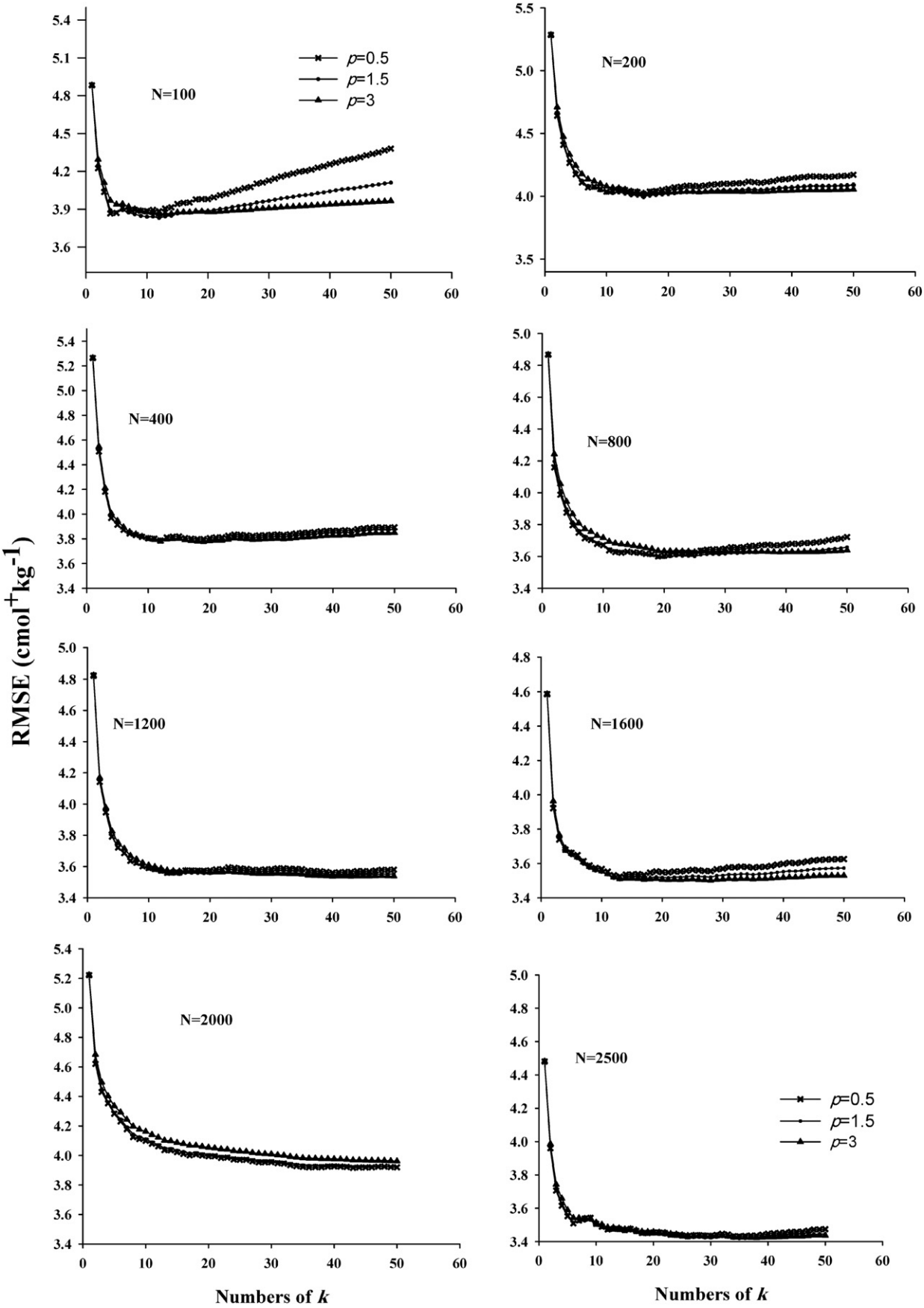

Fig. 5. Variations of the root mean square differences (RMSE) with the number of nearest neighbors (k) as a function of the power term p in Eq. (5).

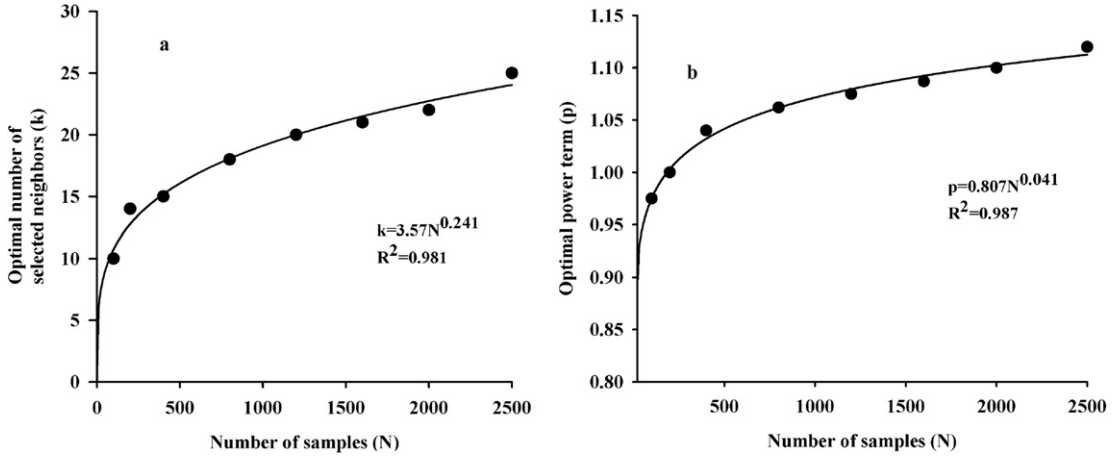

Fig. 6. Effect of data set size on (a) the optimal choice of the number of selected neighbors, and (b) the p term used in weighing the selected neighbors.

Table 3
Comparison of accuracy of the *k*-nearest neighbor technique with optimized *p* and *k* settings calculation using cross validation and Eqs. (12) and (13).

	Evaluation criteria	Cluster 1	Cluster 2	Cluster 3
Size of cluster		620	1350	1400
Optimum *p* and *k*	RMSE	3.70	3.37	3.62
using cross validation	MR	0.08	0.30	0.28
Optimum *P* and *k*	RMSE	3.75	3.38	3.66
using Eqs. (12) and (13)	MR	0.09	0.28	0.28

reason, we also ran a similar analysis for the reference database sizes, 100, 200, 400, 800, 1200, 1600, 2000 and 2500. The optimal value of *k* and *p* in each of the data sets were obtained such that their RMSE values were the minimum. The *k* and *p* values of the best models were shown for each reference data set size in Fig. 6. An increasing trend with increasing data set size were found and the best-fitting equation relating the *k* and *p* values to the reference data set size N was derived based on a power function as:

$$k_{opt} = 3.57N^{0.241} \quad (12)$$

$$p_{opt} = 0.807N^{0.041} \quad (13)$$

where k_{opt} and p_{opt} are best p and *k* values.

In this paper, the power functions were presented describing the relationship between the reference database size and the parameters of *k*-NN method. Lall and Sharma (1996) suggested that the optimal value of *k* is approximately the square root of the number of samples. However, Nemes et al. (2006) presented a power function ($k = 0.655N^{0.493}$) to determine the optimal value of *k* for the prediction of soil water content in 33 and 1500 kPa soil suction. Botula et al. (2012) found also a power function ($k = 0.724N^{0.468}$) to determine the optimum value of *k*. The optimal values of *k* calculated in this study were consistent with the results of Nemes et al. (2006) and Botula et al. (2012) but not with those of Lall and Sharma (1996).

A smaller *k* in a smaller size of reference data set implied preference of the algorithm to use smaller and more meaningful data instead of using a wide range of information (Nemes et al., 2006). It also reduced its impact to modify the optimal value *p*. A smaller *p* also mean that less weight should be considered for those soils at a closer distance than the soils at a greater distance to the nearest neighbor devoted for. In practical terms from a smaller data set, fewer instances were selected, but they were balanced more equally and the opposite is true for larger data sets. The results showed that the selection of a soil as the nearest neighbor ($k = 1$) may cause excessive and unreasonable error in estimating CEC. Thus, choosing $k = 1$ should be avoided.

The empirical equations that are given in this paper were found to determine the optimum values of *k* and *p*. These equations were suitable for our data set and may not be suitable when the approach is applied for another data set. For this reason, the following test was performed to check the validity of Eqs. (12) and (13) to determine the optimal *k* and *p* values.

First, the entire data set were divided into 3 clusters on the basis of soil characteristics and fuzzy k-means clustering (Triantafilis et al., 2003). Then the 20% of the data in each cluster were used for testing, and the other 80% saved as a reference data set. In each cluster the optimal values of *k* and *p* were determined such that the value of RMSE was the lowest. Also the values of *k* and *p* were estimated using Eqs. (12) and (13) and the program was re-run using the new values obtained by the above mentioned procedure and the RMSE and MR of the model were once more calculated.

Table 3 shows the obtained values of RMSE and MR. The results showed that the proposed equations to determine *k* and *p* provide acceptable results. Although the use of Eqs. (12) and (13) can possibly slightly increase error, but the error is statistically not significant. The maximum differences between the RMSE and MR of two methods for determining the optimal values of *k* and *p* were 0.05 and 0.02 $cmol^+ kg^{-1}$, respectively. The findings suggest that these

Table 4
Summary of results, in terms of root-mean-squared error (in $cmol^+ kg^{-1}$), for the *k*-nearest neighbor technique with optimized settings and the neural network models. (SSC, sand, silt and clay content; OC, organic carbon content).

		2500		2000		1600		1200		800		400		200		100	
Estimated method	Input attributes	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
k-NN	SSC + OC	3.63^{ab}	0.17	3.59^{a}	0.23	3.56^{a}	0.17	3.57^{a}	0.077	3.81^{bc}	0.203	3.85^{c}	0.10	3.88^{c}	0.24	3.94^{c}	0.17
	OC + clay	3.62^{a}	0.13	3.62^{a}	0.16	3.59^{a}	0.12	3.58^{a}	0.11	3.79^{b}	0.19	3.81^{b}	0.11	3.91^{b}	0.23	3.96^{b}	0.11
ANN	SSC + OC	3.53^{ab}	0.08	3.56^{abc}	0.09	3.48^{a}	0.04	3.52^{ab}	0.07	3.60^{bc}	0.09	3.63^{c}	0.11	3.75^{d}	0.08	3.76^{d}	0.12
	OC + clay	3.58^{ab}	0.07	3.60^{abc}	0.06	3.54^{a}	0.05	3.58^{ab}	0.09	3.64^{bc}	0.07	3.66^{bc}	0.10	3.70^{c}	0.10	3.70^{c}	0.13

$^{a-d}$ shows the significant differences among RMSE values in each row.

Table 5
Summary of results, in terms of mean residuals (in $cmol^+ kg^{-1}$), for the *k*-nearest neighbor technique with optimized settings and the neural network models. (SSC, sand, silt and clay content; OC, organic carbon content).

		2500		2000		1600		1200		800		400		200		100	
Estimated method	Input attributes	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
k-NN	SSC + OC	0.01	0.31	−0.02	0.25	−0.032	0.23	0.14	0.26	0.05	0.33	−0.33	0.45	−0.05	0.30	−0.18	0.46
	OC + clay	0.10	0.14	0.12	0.09	0.15	0.13	0.11	0.18	0.13	0.35	0.26	0.29	0.10	0.36	0.17	0.43
ANN	SSC + OC	−0.03	0.14	0.02	0.18	0.08	0.07	0.005	0.19	0.004	0.25	0.05	0.20	−0.22	0.32	0.053	0.42
	OC + clay	−0.01	0.12	0.05	0.14	0.03	0.06	−0.1	0.27	0.03	0.21	0.04	0.30	−0.24	0.21	0.08	0.17

Table 6
Correlations (R^2) between estimation errors and the various input attributes of the models that were developed from the data sets with 2500 soils, using soil texture, organic matter content and fractal dimension as input.

	Sand	Silt	Clay	OC
k-NN	0.0020	0.0003	0.0024	0.0029
ANN	0.0005	0.0001	0.0008	0.00003

Eqs. 12 and 13 can be used for determining the optimal *k* and *p* to estimate the CEC in our data set.

3.2.2. Comparison of k-NN and ANN Models

After determining the optimum values of *p* and *k*, *k*-NN approach was run for the inputs and 8 different data set sizes. Also the ANN model was performed to the same data input. On the both modeling methods, the higher number of input variables can relatively improve the estimation of CEC. But this improvement was not statistically significant at the 0.05 level. Furthermore, our results show that increasing the size of the reference data set to a certain amount up to N = 1200 have resulted in a significant reduction in prediction RMSE. However, no significant difference between the accuracy of *k*-NN and ANN methods were detected for N > 1200 (Tables 4 and 5). These findings show that for users with a small data set, the loss in estimated performance by using a similar data range does not seem to be significantly larger than the loss with the ANN technique in the same situation. In most of the cases, the average RMSE of ANN models were smaller than the *k*-NN models. The maximum difference in term of RMSE between the ANN and *k*-NN model was 0.28 $cmol^+ kg^{-1}$. An independent one-sample t-test was run and evaluated at the 0.05 significance level and the results indicated that the RMSE values generated by the ANN and *k*-NN models were statistically different in the case of a reference data set with smaller sizes (N < 1200). But the results showed no significant difference between ANN and *k*-NN models in the prediction of CEC in reference data set when N was greater than 1200.

The bias in estimating CEC were further evaluated (Table 5). The largest unbiased estimate of CEC using ANN and *k*-NN methods were 0.39 and −0.33 $cmol^+ kg^{-1}$, respectively. The bias in *k*-NN approach was positive in most of the estimates. These results indicate that the *k*-NN method underestimates the CEC values of the soil. Nemes et al. (2006) also showed that the *k*-NN method estimated soil moisture content less than the measured value. The predictive ability of the *k*-NN and ANN models in terms of the bias (MR) and overall error (RMSE) don't depend on the combination of input attributes (Tables 4 and 5). The estimation quality is not significantly different when one set of input attributes were used instead of another set. For example, the use of OC and clay did not reduce considerably the quality of the prediction of CEC. Other reports such as by Botula et al. (2013), rather indicated that the accuracy of *k*-NN algorithm in predicting water retention was dependent on the combined effect of the input attributes.

To further investigate, the correlations between estimation errors and the input variables of the models were examined attempting to reveal any systematic distribution of the estimation errors along any of the input variables that were used (Table 6). These correlations were obtained using 2500 samples in the reference data set with all the variables used in the model. The results were shown in terms of R^2 of the linear regression between all the data pairs. In ANN model, R^2 always remained smaller than 0.0008, which indicated that the errors were independent of the model input. The *k*-NN technique showed a higher R^2 values at ≤0.0029. The highest value of R^2 was observed between the estimated error and OC, but the R^2 value was still small enough (0.0029). Similar results were obtained for other reference database sizes.

Sensitivity analysis of *k*-NN model showed that clay content and organic matter can explain 30% and 28% of CEC variance respectively (Fig. 7). Peinemann et al. (2000) and MacDonald (1998) showed that clay content and organic matter can also explain 29 and 20% of CEC variance, respectively.

ANN model sensitivity analysis showed that clay content was the most important variable in estimating CEC and it can explain 29% of the changes in CEC, while the OC and sand content can explain 28 and

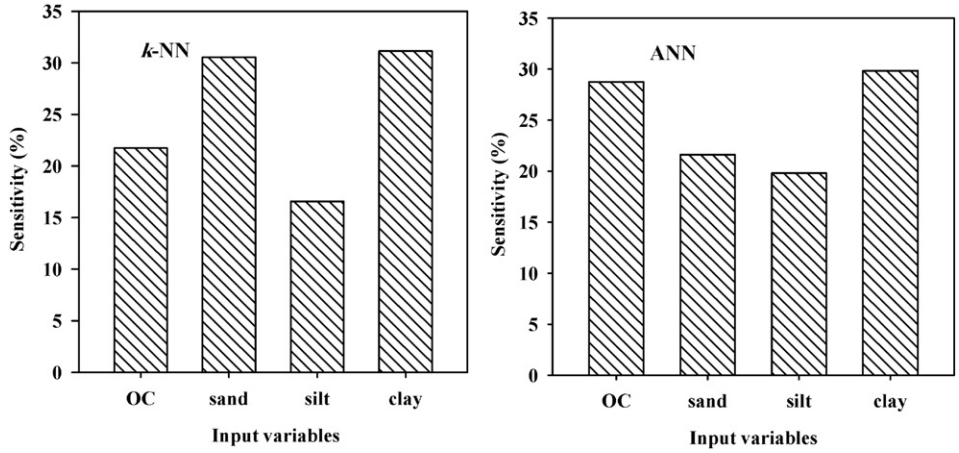

Fig 7. Sensitivity analysis of output variable (CEC) to the input attributes of the models.

23% of CEC variation (Fig. 7) respectively. Bayat et al. (2014) reported that OC can explain 19% CEC variation using ANN model.

4. Conclusion

In this study, we adopted the *k*-NN algorithm developed by Nemes et al. (2006) as a PTF to predict CEC from more easily measured soil properties. This approach was non-parametric; it takes two parameters that should be optimized before being implemented. These parameters included the number of nearest neighbors represented as *k* and the weighing between selected nearest neighbors represented by the parameter *p*. Our results showed that the algorithm efficiency was not dependent on these two parameters. The overall prediction performance of the non-parametric *k*-NN approach was compared with ANN models. It was found that *k*-NN approach could well compete with many other methods of pedotransfer functions (PTFs) because the results showed no significant difference between *k*-NN approach and ANN models. Similarly, the presented *k*-NN variant provides a great degree of flexibility and extra options to the user. The user can, for example, (i) incorporate additional data by appending to or replacing the reference database without the need for developing new equations; (ii) develop the estimations in real time, decide in real time what inputs to use, and may change them from sample to sample if desired. There is also a room for the user to be able to improve a specific local data with a locally available data in a reference database without any significant effects on other available parts of a data in the reference database. For future research, we recommend testing the ability of this technique to predict the CEC of other soils found in the arid and semiarid regions.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments and suggestions that improved the quality of the paper.

References

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. 53, 748–757.

Bannayan, M., Hoogenboom, G., 2009. Using pattern recognition for estimating cultivar coefficients of a crop simulation model. Field Crop Res. 111, 290–302.

Bayat, H., Davatgar, N., Jalali, M., 2014. Prediction of CEC using fractal parameters by artificial neural networks. Int. Agrophys. 28, 143–152.

Bell, M.A., van Keulen, H., 1995. Soil pedotransfer functions for four Mexican soils. Soil Sci. Soc. Am. J. 59, 865–871.

Botula, Y.-D., Cornelis, W.M., Baert, G., Van Ranst, E., 2012. Evaluation of pedotransfer functions for predicting water retention of soils in Lower Congo (D.R. Congo). Agric. Agric. Water Manag. 111, 1–10.

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.

Clark, M.P., Gangopadhyay, S., Brandon, D., Werner, K., Hay, L., Rajagopalan, B., Yates, D., 2004. A resampling procedure for generating conditioned daily weather sequences. Water Resour. Res. 40, 1–15.

Confalonieri, R., Bellocchi, G., Bregaglio, S., Donatelli, M., Acutis, M., 2010. Comparison of sensitivity analysis techniques: a case study with the rice model WARM. Ecol. Model. 221, 1897–1906.

Frate, F.D., Ferrazoli, P., Schiavon, G., 2003. Retrieving soil moisture and agricultural variables by microwave radiometry using neural network. Remote Sens. Environ. 84, 174–183.

Gee, G.W., Bauder, J.W., 1986. Particle-Size Analysis. In: Klute, A. (Ed.), Methods of Soil Analysis Part 1. Physical and Mineralogical Methods, 2nd ed. Monograph 9. Soil Sci. Soc. America J., Madison, Wisconsin, pp. 404–408.

Haghverdi, A., Ghahraman, B., Khoshnood Yazdi, A.A., Arabi, Z., 2010. Estimating of water content in FC and PWP in worth and worth east of Iran's soil samples using *k*-nearest neighbor and artificial aeural aetworks. J. Water Soil 24 (4), 804–814 (In Persian).

Hengl, T., Toomanian, N., Reuter, H., Malakouti, M.J., 2007. Methods to interpolate soil categorical variables from profile observations: Lessons from Iran. Geoderma 140, 417–427.

Jagtap, S.S., Lall, U., Jones, J.W., Gijsman, A.J., Ritchie, J.T., 2004. Dynamic nearest-neighbor method for estimating soil water parameters. Trans. ASAE 47, 1437–1444.

Jalali, V.R., Homaee, M., 2011. A nonparametric model by using k-nearest neighbor technique for predicting soil saturated hydraulic conductivity. J. Water Soil 25, 347–355 (In Persian).

Keller, A., von Steiger, B., van der Zee, S.T., Schulin, R., 2001. A stochastic empirical model for regional heavy metal balances in agro-ecosystems. J. Environ. Qual. 30, 1976–1989.

Krogh, L., Madsen, H.B., Greve, M.H., 2000. Cation exchange capacity pedotransfer functions for Danish soils. Acta Agric. Scand. Sect. B Soil Plant Sci. 50, 1–12.

Lall, U., Sharma, A., 1996. A nearest-neighbor bootstrap for resampling hydrologic time series. Water Resour. Res. 32, 679–693.

Lopez, H.F., Ek, A.R., Bauer, M.E., 2001. Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method. Remote Sens. Environ. 77, 251–274.

MacDonald, K., 1998. In: Pachepsky, Y., Rawls, W.J. (Eds.), Development of Pedotransfer Functions of Southern Ontario Soils. Elsevier, Harrow, Ontario, Canada.

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. 50, 787–794.

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

McBratney, A.B., Minasny, B., Cattle, S.R., Vervoort, R.W., 2002. From pedotransfer functions to soil inference systems. Geoderma 109, 41–73.

Nelson, D.W., Sommers, L.E., 1982. Total carbon, organic carbon and organic matter. In: Page, L.A. (Ed.), Methods of Soil Analysis, Part 2, Chemical and Microbiological Properties, 2nd ed. Monograph 9. Soil Sci. Soc. America J., Madison, Wisconsin, pp. 539–579.

Nemes, A., Rawls, W.J., Pachepsky, Y.A., 2006. Use of the nonparametric nearest neighbor approach to estimate soil hydraulic properties. Soil Sci. Soc. Am. J. 70, 327–336.

Nemes, A., Roberts, R.T., Rawls, W.J., Pachepsky, Y.A., van Genuchten, M.T., 2008. Software to estimate −33 and −1500 kPa soil water retention using the non-parametric *k*-nearest neighbor technique. Environ. Model. Softw. 23, 254–255.

Peinemann, N., Amiotti, N.M., Zalba, P., Villamil, M.B., 2000. Effect of clay minerals and organic matter on the cation exchange capacity of silt fractions. J. Plant Nutr. Soil Sci. 163, 47–52.

Seybold, C.A., Grossman, R.B., Reinsch, T.G., 2005. Predicting cation exchange capacity for soil survey using linear models. Soil Sci. Soc. Am. J. 69, 856–863.

Staff, S.S., 1993. Soil survey manual. USDA Handbook No 18. United States Department of Agriculture, Washington, DC.

Triantafilis, J., Odeh, I.O.A., Minasny, B., McBratney, A.B., 2003. Elucidation of physiographic and hydrogeological features of the lower Namoi valley using fuzzy *k*-means classification of EM34 data. Environ. Model. Softw. 18, 667–680.

Yates, D., Gangopadhyay, S., Rajagopalan, B., Strzepek, K., 2003. A technique for generating regional climate scenarios using a nearest-neighbor algorithm. Water Resour. Res. 39, 1114–1121.