Predicting Cationic Exchange Capacity in Calcareous Soils of East-Azerbaijan Province, Northwest Iran

Farrokh Asadzadeh, Mahdi Maleki-Kakelar & Farzin Shabani

To cite this article: Farrokh Asadzadeh, Mahdi Maleki-Kakelar & Farzin Shabani (2019) Predicting Cationic Exchange Capacity in Calcareous Soils of East-Azerbaijan Province, Northwest Iran, Communications in Soil Science and Plant Analysis, 50:9, 1106-1116, DOI: 10.1080/00103624.2019.1604728

To link to this article: https://doi.org/10.1080/00103624.2019.1604728

Published online: 18 Apr 2019.
Predicting Cationic Exchange Capacity in Calcareous Soils of East-Azerbaijan Province, Northwest Iran

Farrokh Asadzadeh*, Mahdi Maleki-Kakelarb, and Farzin Shabani

daDepartment of Soil Science, Faculty of Agriculture, Urmia University, Urmia, Iran; bDepartment of Chemical Engineering, Faculty of Engineering, University of Zanjan, Zanjan, Iran; cARC Centre of Excellence for Australian Biodiversity and Heritage, Global Ecology, College of Science & Engineering, Flinders University, Adelaide, South Australia, Australia; dDepartment of Biological Sciences, Macquarie University, Sydney, New South Wales, Australia

ABSTRACT

The aim of this research is to study the efficiency of pedotransfer functions (PTFs) and artificial neural networks (ANNs) for cationic exchange capacity (CEC) prediction using readily available soil properties. Here, 417 soil samples were collected from the calcareous soils located in East-Azerbaijan province, northwest Iran and readily available soil properties, such as particle size distribution (PSD), organic matter (OM) and calcium carbonate equivalent (CCE), were measured. The entire 417 soil samples were divided into two groups, a training data set (83 soil samples) and test data set (334 soil samples). The performances of several published and derived PTFs and developed neural network algorithms using multilayer perceptron were compared, using a test data set. Results showed that, based on statistics of RMSE and R², PTFs and ANNs had a similar performance, and there was no significant difference in the accuracy of the model results. The result of the sensitivity analysis showed that the ANN models were very sensitive to the clay variable (due to the high variability of the clay). Finally, the models tested in this study could account for 85% of the variations in cationic exchange capacity (CEC) of soils in the studied area.

Abbreviations: ANN: artificial neural networks; MLP: multilayer perceptron; MLR: multiple linear regression; PTFs: Pedotransfer Functions; RBF: Radial Basis Function; MAE: mean absolute error; MSE: mean square error; CEC: cationic exchange capacity

ARTICLE HISTORY

Received 29 January 2018
Accepted 14 March 2019

KEYWORDS

Artificial neural networks; Pedotransfer functions; Readily available soil properties

Introduction

Cationic exchange capacity (CEC) as an important chemical characteristic of soil, represents the soil’s ability to absorb and release plant nutrient requirements and provides a buffer against soil acidification. It may be also used to estimate the potential risk of heavy metals and some organic cationic pollutants (Jones and Jacobsen 2005). Typically, CEC is a reliable index to characterize the quality and productivity of soil and is highly affected by physicochemical factors including the amount of Organic Matter (OM), the percentage and the type of clay and soil conditions (Khaledian et al. 2017b). As a result, determining of the CEC can be of major significance to characterize the soil on the content of ionic elements, clay content, texture, and provide information about the possible requirement for fertilization and neutralization of the soil acidity (Aprile and Lorandi 2012). However, the analytical laboratory procedures for measuring CEC is difficult, time-consuming and costly (Manrique, Jones, and Dyke 1991; Krogh, Breuning-Madsen, and Greve 2000; Ghorbani et al. 2015; Jones 2012; Kovačević, Bajat, and Gajić 2010), so, estimating CEC based on indirect approaches such as pedotransfer functions (PTFs) by using regular soil test results is feasible and satisfying.
In order to predict the CEC in soils, PTFs are suitable that correlate available soil information such as texture, pH, and organic carbon (OC) to CEC for parameterization of soil processes (Borggaard et al. 2004; Reidy et al. 2016). In most studies, CEC is usually considered as a linear function of organic matter (OM) and clay content, and multiple linear regression (MLR) analysis is applied to find the coefficients of the model (Horn, During, and Gath 2005; McBratney et al. 2002; Seybold, Grossman, and Reinsch 2005; Wang, Li, and Klassen 2005).

Recently, artificial neural networks’ (ANNs), as an innovative approach to model PTFs, have been appropriately applied to predict some soil properties that their measurement is challenging (Amini et al. 2005; Bayat, Davatgar, and Jalali 2014; Emamgolizadeh et al. 2015; Minasny and Mcbratney 2002). As a method of intelligent data processing, ANNs are able to calculate continuous nonlinear function with any degree of precision, and so they have been attracted the attention of researchers in recent years (Jain et al. 1999; McBratney et al. 2002; Pham and Karaboga 2012). One advantage of using ANNs compared to the traditional statistical techniques such as regression is that they do not determine necessarily a specific function to express the relationship between the input and output, as the relationship would be achieved through a training process (Rojas 2013). For instance, multilayer perceptron (MLP) method of artificial neural network uses an error back propagation algorithm to develop the Pedotransfer Functions (PTFs) (Kashi, Emamgholizadeh, and Ghorbani 2014; Oh 2011; Vereecken et al. 2010).

Although CEC prediction using statistical methods is a preference in environmental soil science, selecting the most promising methods is still a field that requires additional studies. Furthermore, models developed for one region may not give acceptable predicts for a different region (Amini et al. 2005; Emamgolizadeh et al. 2015; Kalkhajeh et al. 2012; Wagner et al. 2001). Hence, the aim of this study was to develop the PTF and ANN models to estimate CEC in calcareous soils East-Azarbaijan province, north-west Iran, and to compare the performance of the two models in order to detect most productive model.

**Materials and methods**

**Study area and soil sampling**

The present study was conducted in the northwestern Iranian province of East Azerbaijan. The province lies between latitude 36° 45′ to 39° 26′N and longitude 45° 7′ to 48° 20′E, with an area equal to 45,846 square kilometers (approximately 2.81% of the total country area) East Azerbaijan is a mountainous area with 40% of its surface being high mountains, 28.2% foothills, and 31.8% intermountain plains. At the study area, the annual mean precipitation ranges from 361 to 608 mm except in the high mountain areas. The total arable land is about 1.22 million hectares, which is 26.6% of the total surface area of the province. A total of 417 surface soil samples (0–20) were taken from the agricultural lands of the study area (Figure 1).

**Physical and chemical analysis**

Soil samples were air-dried at 15–25°C and then were passed through a 2-mm sieve and this size fraction was used for further laboratory analyses. Distribution of particle size in the soil samples was determined by the hydrometer method (Gee, Bauder, and Klute 1986) and organic carbon content was determined by dichromate oxidation method. Calcium carbonate equivalence (CCE) was determined by the acid neutralization method (Rowell 2014). The cation exchange capacity of the soils was measured using the ammonium acetate method (Swift and Sparks 1996).

**Statistical analysis and data processing**

In this study, initially, data were normalized through Equation 1 to increase the speed and accuracy of the network. Thus, data were converted into numbers between 0 and 1, as the output of most of the functions is in the threshold of 0 to 1.
\[ x_n = \frac{x_{\text{max}} - x_i}{x_{\text{max}} - x_{\text{min}}} \]  

where \( x_{\text{max}} \) is the maximum data, \( x_{\text{min}} \) the minimum data, \( x_i \) observed data and \( x_n \) is the normalized data. One set of data was used to train the neural network and regression models and another set to test the constructed models; 80% and 20% of data were used at random in order to train and test models, respectively.

**Modeling approach and methodology**

The general purpose of multiple regressions is to make predictions based on the relationship between several independent variables and a dependent or criterion variable. Multiple regressions are the most common method used in development PTFs (Seyedmohammad, Esmaelnejad, and Ramezanpour 2016). After normalizing data, multiple linear regression function was derived for the training data set. In this method, all data were first inserted as input data and subsequently, the data that were significantly less effective on output parameter were eliminated. PTFs of regression models for estimating the CEC in this work included (PTFs (1)) (Minasny and Mcbratney 2002; Minasny, McBratney, and Bristow 1999), (PTFs (2)) (McBratney et al. 2002) and derived PTFs in this study (PTFs (3)). In the case of PTF (3), it is assumed that the CEC is a linear function of clay content, sand content, OM and calcium carbonate.

In order to develop models of neural networks, similar to the regression models, variables, which were significantly correlated with CEC, were used as inputs to the networks. These networks were made of three layers: input, hidden and output layers (Figure 2). This kind of network which a large number of simple processing elements that are connected to each other by the weighted connections, according to the required specified architecture, is called multilayer feed forward neural network (FFNN) (Dawson and Wilby 1998). In this study, developed FFNN models were multi-layer perceptron which is the most frequently used neural network structure in ecological modeling and soil science (Agyare, Park, and...
For this purpose, three multilayer perceptrons (MLP), known as the feed-forward neural network (FFNN), with error backpropagation algorithm were used in this study to find the optimum number of inputs. Determination of the number of neurons in the hidden layer of the neural network is an important step and plays a significant role in the performance of the neural network (Khataee and Kasiri 2010). To determine the optimum neural network, various numbers of hidden neurons layer and different types of threshold function in each layer were determined. Where two different networks had the same error, the network having less neurons was selected. Combinations of threshold functions of tansig, logsig (two sigmoid functions) and purelin (linear function) were used as activation functions of developed neural networks. Accordingly, and by considering the performance of the constructed networks, the threshold function of each layer was determined to estimate the CEC.

**Performance criteria**

A series of criteria were used to evaluate the effectiveness of different models in estimating the dependent variable (CEC). These criteria include the coefficient of determination ($R^2$) (Equation 2), mean error (ME) (Equation 3) and root-mean-square error (RMSE) (Equation 4).

$$R^2 = \frac{\sum_{i=1}^{n} (O_i - \bar{O})(P_1 - \bar{P}_1)^2}{\sum_{i=1}^{n} (O_i - \bar{O})^2 \sum_{i=1}^{n} (P_1 - \bar{P}_1)^2}$$  \hspace{1cm} (2)

$$ME = \left(\frac{1}{n}\right) \sum_{i=1}^{n} (O_i - P_i)$$  \hspace{1cm} (3)

$$RMSE = \sqrt{\left(\frac{1}{n}\right) \sum_{i=1}^{n} (O_i - P_i)^2}$$  \hspace{1cm} (4)

where $O_i$ and $P_i$ are observed and predicted CEC, respectively, and $n$ is the number of data. These accuracy indicators were used for model evaluation and selection to estimate soil CEC in which high coefficient of determination values ($R^2$) and low RMSE values reveal the capability of the models. All analysis of issues related to data analysis techniques were performed by the software MATLAB v. 8.5.0.197613.
Results & discussion

Data summary statistics

The statistical description of readily available soil properties and CEC of the samples are presented in Table 1. Among the readily available soil properties, OM of the soil has the lowest coefficient of variation (7.14%). However, the coefficient of variation of CCE is higher than all the other features and is about 83.70% (Table 1). Significant changes in the CCE can be attributed to differences in topography and climatic conditions of the areas under study. The coefficient of variation of the CEC in the area under study was 37.73%.

A study of the correlation coefficient between variables showed that the lowest and the highest correlation coefficient belonged to CCE with \( r = 0.169 \) and clay with \( r = 0.741 \), respectively. Moreover, there was a significant negative correlation \( r = -0.69 \) between the amount of sand in the soil and the CEC content. When the percentage of sand rises, the amount of particles with negative charge drops. The lower CEC content could be explained by the decrease in the amount of particles with negative charge, which leads to a significant decrease in the absorption of cations.

CEC is extremely influenced by soil physical (e.g., soil texture), chemical (e.g., pH, mineralogy), and biological (e.g., OC) properties. Accordingly, the relationship between CEC and soil properties that are readily available is highly significant (Khaledian et al. 2017a, 2017b). For example, it has been reported that CEC had a significant negative correlation with sand content (Shabani and Norouzi 2015). However, this relationship is not always true. Amini et al. (2005) observed that sand and silt had an insignificant effect in predicting CEC. In another study, Ghorbani et al. (2015) found that silt has a lower effect on the prediction of soil CEC than clay and sand. Furthermore, according to Zeraatpishe and Khormali (2012) high amounts of OM can affect soil pH and, consequently, CEC (Zeraatpishe and Khormali 2012). Although many studies indicate that CEC is governed by OM content (McBratney et al. 2002; Ulusoy et al. 2016), the results in this study did not support that. Generally, CEC in this study, positive correlation was observed between CEC and clay whereas CEC was negatively correlated to sand and pH. It may be that the sampling depth has an important impact on this. There are numerous studies that assay CEC only sample to about 0–20 cm depth (Khaledian et al. 2017a; Kweon, Lund, and Maxton 2013). Soil OM not only tends to be accumulated in these shallow layers but also changes substantially from place to place even in surface soils (Parras-Alcántara et al. 2015).

Soil texture distribution is shown in Figure 3. As Figure 3 shows clearly, the studied soils constituted many of texture classes (9 of 12 texture classes).

The results of applying the PTFs are shown in Table 2. Comparing the PTFs indicates, that based on RMSE and \( R^2 \) statistics, these models have the same performance. The estimated and measured CEC for different PTFs is shown in Figure 4. These results show that the reviewed and devised PTFs could account 82 to 83% of the CEC variation in the studied region.

In all three models, the activation functions from the input layer to the hidden sigmoid function (tansig and logsig) and the hidden layer to the output linear function (purelin) were selected (Table 3). In the first model (ANN1), only organic carbon and clay were used as inputs. In the second model (ANN2), clay percentage and multiplication of clay in organic carbon were used as inputs. Finally, in the third model (ANN3), four inputs including clay, sand, OM and CCE were used as input variables.

<table>
<thead>
<tr>
<th>Readily available properties</th>
<th>CV (percentage)</th>
<th>Standard deviation</th>
<th>Mean</th>
<th>Max</th>
<th>Min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sand</td>
<td>38.94</td>
<td>20.32</td>
<td>52.16</td>
<td>96.00</td>
<td>6.00</td>
</tr>
<tr>
<td>Silt</td>
<td>40.75</td>
<td>13.30</td>
<td>32.60</td>
<td>70.00</td>
<td>3.20</td>
</tr>
<tr>
<td>Clay</td>
<td>65.06</td>
<td>9.84</td>
<td>15.10</td>
<td>46.00</td>
<td>6.00</td>
</tr>
<tr>
<td>OM</td>
<td>7.14</td>
<td>0.15</td>
<td>2.01</td>
<td>2.65</td>
<td>1.59</td>
</tr>
<tr>
<td>CCE</td>
<td>83.70</td>
<td>8.72</td>
<td>10.41</td>
<td>80.89</td>
<td>0.00</td>
</tr>
<tr>
<td>CEC</td>
<td>37.73</td>
<td>7.96</td>
<td>20.58</td>
<td>38.65</td>
<td>6.54</td>
</tr>
</tbody>
</table>
ANN1, ANN2, and ANN3 with 10, 8, and 3 neurons in the hidden layer, respectively, had the highest accuracy and lowest error. The results of selected ANN models are shown in Table 3. Moreover, the scatter plot of the measured against predicted CEC for the test data set is given in Figure 5 for the selected ANN models.

In order to investigate the effect of each of the input variables on the CEC, it is necessary to apply a sensitivity analysis. In this study, the method of eliminating variables and their impact on increasing the amount of RMSE was used. That variable whole deletion having the largest effect on increasing the RMSE was considered as the most sensitive variable. Since model ANN2 had the highest R² and the lowest RMSE sensitivity, an analysis was conducted for this model. The results of the sensitivity analysis are presented in Table 4. In the table, grade 1 is assigned to the variable whose deletion can produce the highest amount of RMSE and grade 2 to the OM with less variability in the study area.

In order to evaluate the efficiency of the models and compare the ANN and PTF models, test data was used to predict the CEC. As Tables 2 and 3 shows, the coefficient of determination for each of the three ANNs models of (ANN1, ANN2, and ANN3) is slightly greater than the regression models (PTF (1), PTF (2) and is similar in PTF (3)). The amount of RMSE and R² of the CEC estimated by the neural network model ANN2 (where the four inputs of organic carbon, clay content and multiplying the amount of clay were used), compared with other models, was the lowest and greatest, respectively, while the highest and lowest of these statistics belonged to regression model PTFs (3).

![Figure 3. Soil texture classes in the studied soils.](image)

<table>
<thead>
<tr>
<th>Model type</th>
<th>Pedotransfer function</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTFs(1)</td>
<td>CEC = 0.156 + 0.858 Clay + 0.177 OM</td>
<td>ME -0.124</td>
<td>RMSE 3.15</td>
</tr>
<tr>
<td>PTFs(2)</td>
<td>CEC = 0.112 + 0.878 Clay + 0.127 ClayOM</td>
<td>ME -0.126</td>
<td>RMSE 3.13</td>
</tr>
<tr>
<td>PTFs(3)</td>
<td>CEC = 0.027 + 0.811 Clay − 0.045 Sand + 0.168 OM − 0.091 CaCO3</td>
<td>ME -0.123</td>
<td>RMSE 3.12</td>
</tr>
</tbody>
</table>
Several factors affect the performance of the models, including the differences in the morphology of sampling areas, and variation of soil forming factors in the study areas. Another important factor affecting the performance of neural networks and regression models is the kind of relationship between the dependent and independent variables. The linear relationship between the independent and objective variables in this study is one of the reasons for the similarity of MLR compared to ANN models. Amini et al. (2005) indicated that there were no significant differences between the neural network models and the regression models for predicting CEC (Amini et al. 2005). The results of this study did not match with those of Kashi, Emamgholizadeh, and Ghorbani (2014) in estimating the CEC the North Iran (Semnan Province) (Kashi, Emamgholizadeh, and Ghorbani 2014). Utilizing different neural networks on 200 soil samples in the Goosheh region, Iran, they showed that the MLP network had a better performance when compared to the neuro-fuzzy network, RBF neural network and multiple regression for prediction of CEC (Kashi, Emamgholizadeh, and Ghorbani 2014). Similarly, Ghorbani et al. (2015) utilized neural networks, neuro-fuzzy and multiple linear regression on 220 soil samples to predict CEC and found that

![Figure 4. Observed versus estimated CEC by PTFs(1), PTFs(2) and PTFs(3) models.](image)

<table>
<thead>
<tr>
<th>Network type</th>
<th>Selected architecture</th>
<th>Transfer functions</th>
<th>Train ME</th>
<th>RMSE</th>
<th>R²</th>
<th>Test ME</th>
<th>RMSE</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN₁</td>
<td>3-10-1</td>
<td>Tansig-purelin</td>
<td>−0.172</td>
<td>2.95</td>
<td>0.86</td>
<td>0.500</td>
<td>2.84</td>
<td>0.84</td>
</tr>
<tr>
<td>ANN₂</td>
<td>4-8-1</td>
<td>Tansig-purelin</td>
<td>−0.027</td>
<td>2.95</td>
<td>0.86</td>
<td>0.620</td>
<td>2.81</td>
<td>0.85</td>
</tr>
<tr>
<td>ANN₃</td>
<td>5-3-1</td>
<td>Tansig-purelin</td>
<td>−0.136</td>
<td>2.93</td>
<td>0.86</td>
<td>0.540</td>
<td>2.83</td>
<td>0.83</td>
</tr>
</tbody>
</table>
neuro-fuzzy networks had better performance than neural networks and other regression models (Ghorbani et al. 2015). Alternatively, several researchers showed that neural networks have higher efficiency compared with the regression models. Using neural networks and regression models on 230 soil samples, Pachepsky, Timlin, and Varallyay (1996) showed that neural network models predicted soil water holding capacity with greater accuracy and lower error than other modeling methods (Pachepsky, Timlin, and Varallyay 1996). Amini et al. (2005) showed that neural network-based models provided more reliable predictions than the regression-based Pedotransfer Functions (PTFs) for the Aridisols of the Isfahan region (Amini et al. 2005). Tang et al. (2009) utilized Radial Basis Function (RBF) networks to predict CEC and found that the RBF model obtained a greater precision than the MLR model (Tang et al. 2009). The results of this study indicated that according to the statistics of the coefficient of determination, there was a small difference between the neural networks of the perceptron (MLP) type and multivariate linear regression test phase (the coefficient of determination was 0.78 for MLR model, and 0.80 for neural network of MLP type). Accordingly, although neural networks with different architectures had a slightly better performance than regression models based on RMSE statistics, it seems that the estimating CEC based on derived

Figure 5. Observed versus estimated CEC, using selected type ANN$_1$, ANN$_2$, and ANN$_3$ models.

| Table 4. Result of sensitivity analysis and determination sensitivity degree of input variables. |
|------------------|------------------|-------|------------------|
| Input variables  | Removed variables | RMSE  | Sensitive degree |
| Clay, Clay$^*$.OM| -                | 2.81  | -                |
| Clay             | OM               | 3.51  | 2                |
| OM               | Clay             | 5.12  | 1                |

COMMUNICATIONS IN SOIL SCIENCE AND PLANT ANALYSIS
PTFs by using MLR can be practically feasible due to its simplicity. In addition, MLR easily provides a way of adjusting for potentially confounding variables that have been included in the model, which would be essential to utilize these results to predict CEC in other locations with acceptable accuracy.

**Conclusion**

As CEC is a useful indicator of soil fertility and is a soil property which is relatively difficult, time-consuming to determine, it is a foremost soil property to be estimated based on indirect approaches. Multiple linear regression and neural network models were developed to forecast CEC, from readily available soil properties in calcareous soils of East-Azerbaijan province, northwest Iran. The performance of ANNs and derived PTF models was assessed using test data sets. Although neural networks with different architectures had a slightly better performance than regression models based on RMSE statistics, it seems that the estimating CEC based on derived PTFs by using MLR can be practically feasible due to its simplicity and generality. Remarkably, in spite of relatively acceptable accuracy in CEC prediction, it can be concluded that further research is needed to achieve a more comprehensive CEC predicting model for a broader range of soils in different geographical regions owing to the limited database in this work.

**Nomenclature**

- $O_i$: observed CEC
- $P_i$: predicted CEC
- $n$: number of data
- $x_{\text{max}}$: maximum data
- $x_{\text{min}}$: minimum data
- $x_i$: observed data
- $x_n$: normalized data

**ORCID**

Farzin Shabani  [http://orcid.org/0000-0002-5100-8921](http://orcid.org/0000-0002-5100-8921)

**References**


