THE APPLICATION OF ARTIFICIAL INTELLIGENCE METHODS FOR EVALUATION AND MODELLING OF THE SOIL PROPERTIES

Vahid PourAmin[1]
Elham Alighardash[2]

1. Sayyed Jamaleddin Asadabadi University, Asadabad, Iran. E-mail: v.pouramin@sjau.ac.ir
2. Sayyed Jamaleddin Asadabadi University, Asadabad, Iran. E-mail: e.alighardash@sjau.ac.ir

Abstract: As artificial intelligence (AI) applications see wider deployment, it becomes increasingly important to study the social and societal implications of AI adoption. AI has been in existence for over six decades and has experienced AI winters and springs. The rise of super computing power and Big Data technologies appear to have empowered AI in recent years. The new generation of AI is rapidly expanding and has again become an attractive topic for research. The paper first provides a view of the history of AI through the relevant papers published in the International Journals. Examination of the soil properties like Cation Exchange Capacity (CEC) plays an important role in the study of environmental researches. The spatial and temporal variability of this property has been caused to the development of indirect methods in estimation of the soil characteristics. This paper aims to employ different AI-based methods to estimate the cation exchange capacity. One hundred and fifty soil samples are collected from different horizons of soil profiles located in the Behbahan region, Khuzestan Province, Southwest of Iran. Finally, multiple linear regression, Neuro-Fuzzy, feedforward back-propagation network, and other methods are employed to develop a pedotransfer function (PTF) for predicting soil parameters using easily measurable characteristics of clay and organic carbon. As an interesting consequence, Neuro-Fuzzy, SVM, and some others are superior to artificial neural networks and multiple linear regression in predicting soil property.

Keywords: Artificial intelligence, neural network, Cation Exchange Capacity, Pedotransfer function.

1. INTRODUCTION

The emergence of artificial intelligence in recent years has garnered widespread controversy. This can be planning to allow us to try to do what the human state is in a position to try to do. Bill Gates mentioned, "humans should be worried about the threat posed by Artificial Intelligence" [1]. Stephen William Hawking, in other words, stated that "the development of complete artificial intelligence can spell the tip of human generation" [2]. These very different opinions invoke leading experts to try to more research on how humans coexist with computer science and the way to reduce the negative impact of technology and benefit from the positive impression of it. There is no standard definition of AI. Commonly named as a machine's ability to be told experiences, accommodates new entrances and do human-like tasks.

With the rapid advancement of big data technologies, as an example, improved computing storage and ultra-fast processing speed, artificial intelligence is being restored with the supply and power of big data. Therefore, after years of hope and promise, AI is gaining meaningful traction within top corporations [3]. It's been reported that the adoption of AI-enabled systems in organizations is expanding rapidly [4], and AI is changing business [5]. The new wave of AI systems predicts the flexibility of a company to use data for forecasting. It has significantly reduced the cost of predicting [6]. In line with Gartner Technology Review 2018 [7], AI has been introduced because of the favored strategic technology. The ability to use AI to bolster decision-making, reinvent business models and ecosystems, and reimburse the customer experience, pays off digital innovation by 2025. Gartner polls show that 59% of organizations are still gathering information to make their own AI strategies. Efforts to make human intelligence like the advances in algorithms and computers have advanced dramatically in recent decades. However, engineering whole human intelligence has been a difficult task. Instead, progress has been made within the engineering of specific human capabilities. While we frequently use the term artificial intelligence today to confer with machine learning, the meaning of artificial intelligence has fluctuated over the past 60 years to put variable emphasis on vision, language, speech, and pattern recognition.

Although early AI research was inspired by several other fields, including some social sciences, modern AI research is increasingly focused on engineering applications—perhaps because of the increasingly central role of the technology industry. Furthermore, the foremost central research institutions within the AI research community are increasingly based in industry instead of academia. This study aims to predict cation exchange capacity (CEC) as one of the most critical attributes of environmental research, by the use of AI methods. The amount of negative and positive change in the soil is known as CEC. It's accessible to bind charged ions (cations). Essential plant nutrients and harmful elements are cations. Cation exchange capacity is employed as a measure of fecundity, nutrient retention capacity, and also the capacity to shield groundwater from cation contamination. CEC delimiters fluctuations in nutrient availability and soil pH. Soil components known to contribute to CEC are clay and organic matter and, to a lesser extent, silt [8].

Tamari et al. [9] gave a review on ANN and their application in predicting soil hydraulic properties. Most researchers have understood that ANN performs better than multiple regressions. Amini et al. [10] tested several published PTFs and developed two neural network algorithms using multilayer perception and general regression neural networks supported a group of 170 soil samples for predicting of cation exchange capacity in central Iran. They found that the neural network-based models provided more reliable predictions than the regression-based PTFs. Schaap et al. [11] used ANNs for predicting some soil hydraulic properties. They also confirmed the applicability of ANNs and concluded that the accuracy of those models depends upon the number of inputs. The target of this study is to estimate the final applicability of artificial neural network, Neuro-Fuzzy and, multivariate regression in determining cation exchange capacity within the soils of Iran.

Akbarzadeh et al. [12] in 2009 suggested using characteristics of the soil to estimate CEC that are more convenient to measure. They collected eighty soil samples were from different horizons of 26 soil profiles. Their selected variables included soil texture, organic carbon, and CEC. MLR, Neuro-Fuzzy, and feedforward back-propagation network were models that utilized to develop a pedotransfer function to predict soil parameters. Results showed that Neuro-Fuzzy was superior to ANN and MLR in predicting soil property.

Keshavarzi and Sarmadian [13] went the better road. They ameliorated the performance of MLR and ANN model for predicting soil parameters such as CEC by deriving a benefit of easily measurable characteristics of clay and organic carbon. The value of RMSE and R^2 gained by the ANN model for CEC were 0.47 and 0.94, while these criteria for the MLR model were 0.65 and 0.88, respectively.

Shekofteh et al. [14] proposed the support vector regression (SVR) combined with genetic algorithm (GA) together with the adaptive network-based fuzzy inference system (ANFIS) to predict soil CEC based on 104 soil samples collected from soil surface under four different land uses. Their results of sensitivity analysis showed that some parameters such as soil organic matter and clay content are more effective than others on both models. The performance of both models was acceptable, but SVR was better than ANFIS. This suggests that SVR and ANFIS are robust tools for the design of PTFs in order to CEC prediction.

Shekofteh et al. [15], in another study, introduced a hybrid algorithm: an advance ant colony organization (ACO) in combination with an adaptive network-based fuzzy inference system (ANFIS). Their approach has to phase of properties or feature selection that influence soil CEC and predict it. They showed that the ANFIS method resulted in higher model efficiency and coefficient of determination (R^2 = 0.91) than multiple linear regression (MLR) approach (R^2 = 0.74).

In the rest of the paper, first materials and methods of doing this study, including data collection methods, features that were selected, and models, are being discussed. Then evaluation measures introduce. At last, results and conclusion investigate.

2. MATERIALS AND METHODS

2.1. DATA COLLECTION AND SOIL SAMPLE ANALYSIS

The study area is found within the Southeast direction of Behbahan city and by the side of the Persian Gulf in Khuzestan province of Iran. This study administered in a region including 150 points. Figure 1 illustrates the placement of the study area.

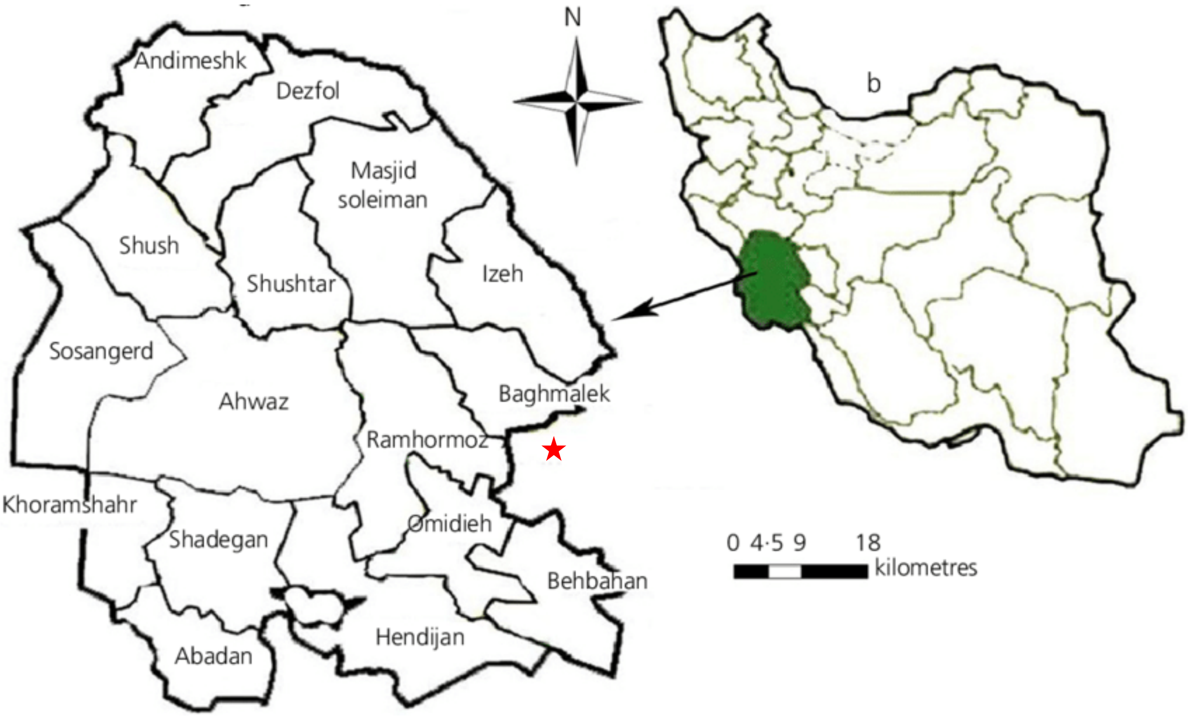

Figure 1: The placement of the study area

After the interpretation of aerial photographs, the digging site of the profile is identified. Some pedons are chosen, so one hundred and fifty soil samples are collected from different horizons of those profiles.

2.2. FEATURE SELECTION

The details of soil properties are shown in Table 1. Electrical conductivity and pH were measured on fresh soil samples in deionized water (soil solution, 1:2.5). Organic Matter (OM) was determined by the potassium dichromate wet combustion procedure.

Table 1: The properties of soil samples

Soil attributes	Min	Max	Average	SD
pH	6.7	7.6	7.13	0.2
Clay	16	70	42.12	10.15
OM	0.31	1.58	0.71	0.24

Min- minimum, Max-maximum, SD-standard deviation

As shekofteh et al. mentioned in their study [15], one of the parameters that effects the result of prediction are features that selected as the input of AI systems. They proved that in CEC prediction, the best composition of features is pH, clay, silt, BD, and OM. Comparison of different inputs and their role in performance showed in table 2.

Table 2: Comparison of ANFIS model performance based on different selected features sets [15]

Number of features	Best feature	RMSE (%)
1	Porosity	3.29
2	OM, BD	2.60
3	Sand, Clay, OM	2.62
4	Silt, PD, Porosity, OM	2.59
5	pH, Clay, Silt, BD, OM	2.53
6	Silt, CCE, BD, PD, Porosiry, OM	2.55
7	pH, Sand, Silt, BD, PD, Porosiry, OM	2.62
8	EC, pH, Sand, Silt, BD, PD, Porosiry, OM	2.68
9	EC, pH, Sand, Silt, Clay, BD, PD, Porosiry, OM	2.96
10	All features	3.16

In this table, SOM or OM is soil organic matter content, CCE stands for calcium carbonate equivalent, BD is equal to bulk density, PD corresponds with particle density, and EC is electrical conductivity. Among the best features, pH, clay, and OM were available and used for this study.

2.3. AI METHODS

Machine learning models as the main domain of AI methods can be categorized into four classes, including neuron-based (MLP, GRNN, and ANFIS), kernel-based (SVM, KNEA), tree-based (M5Tree, XGBoost) and curve-based (MARS) models. The MLP (Multilayer perceptron neural networks) model is one of the common applied ANNs models, covering a feedforward neural network for nonlinear function approximation. The structure of the MLP model consists of layers and neurons. The overall view of this framework is built by the input, hidden, and output layers. The hidden layer can have only one layer or more, but the experiences showed that increasing the number of hidden layers did not affect so much on the performance. Therefore, the number of neurons that are critical must be determined by the trial and error approach. A typical MLP model may be a three-layer neural network. The primary layer was the input layer that is fed by inputs, so its neurons number is equal to the input variables of the problem. The output layer is responsible for generating the answer. In this particular application that the problem is to predicate one variable, the number of output neurons should be one. The MLP model is trained by different learning algorithms, such as the Levenberg–Marquardt algorithm, which interpolates between the Gauss-Newton algorithm (GNA) and the gradient descent algorithm. It's more robust but still can stick with local minima.

The GRNN (Generalized regression neural network) model is proposed by [16] and is one of the radial basis function neural network (RBF) models. GRNN does not require an iterative training procedure the same as back-propagation. This model is an approximator of the nonlinear function of the input and output vectors. The estimation obtained from the training dataset. It shows a parallel structure in the learning process. The ANFIS (Adaptive neuro-fuzzy inference system) model is proposed by [17], which is a composition of multi-layer adaptive neural network with the fuzzy inference system. This model is capable of approximating nonlinear functions with different fuzzy models such as the first-order Sugeno with some fuzzy if-then rules. The ANFIS model is created of five parts: the fuzzification, product, normalization, de-fuzzification, and output unit. The activation, approximation, and learning functions could be diverse. Still, all of them are able to adjust the parameters in a fuzzy inference system, where the forward and backward passes are exploited to reduce the computed errors. A general structure of a fuzzy system is demonstrated in Figure 2.

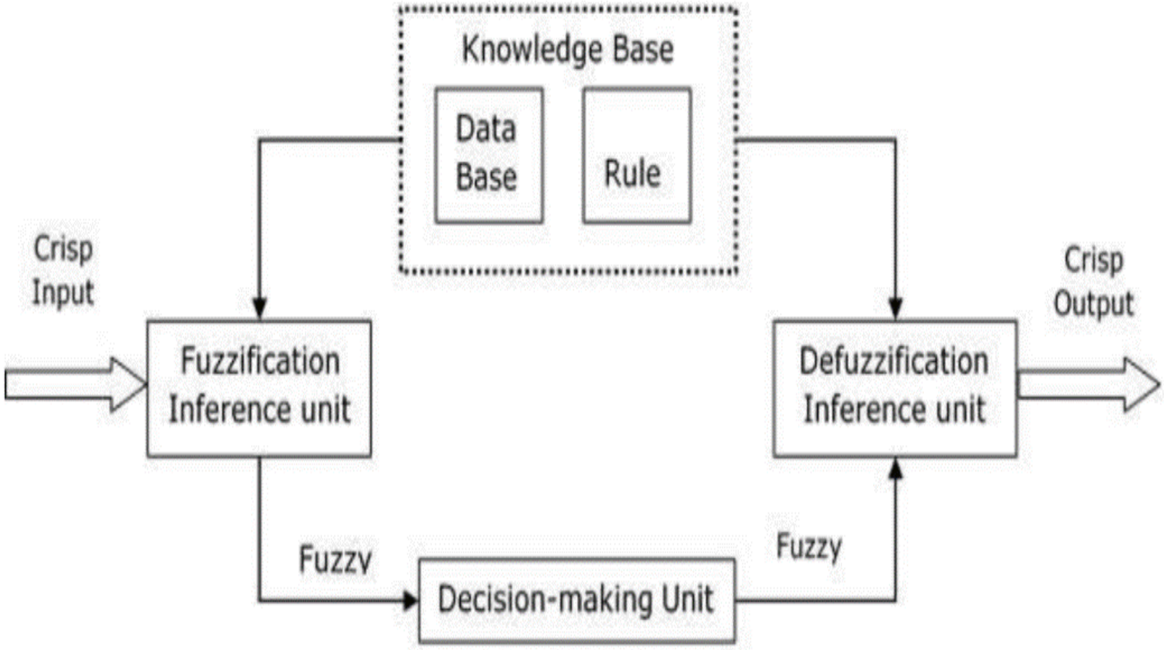

Figure 2: A Fuzzy System Architecture

The corresponding ANFIS architecture is shown in Figure 3. Nodes at the same layer have similar functions. The SVM (Support vector machine) model is developed by [18]. This method has shown excellent results for classification, pattern recognition, and regression analysis. The fundamental part of this model is kernel functions, which transform original, lower-dimensional input dataset to a higher-dimensional feature space implicitly. The SVM model has been successfully applied in predicting models [19, 20]. The radial basis function (RBF) as a nonlinear kernel function is one of the possible options that is selected to use in the present study as a result of its outstanding performance for predicting CEC.

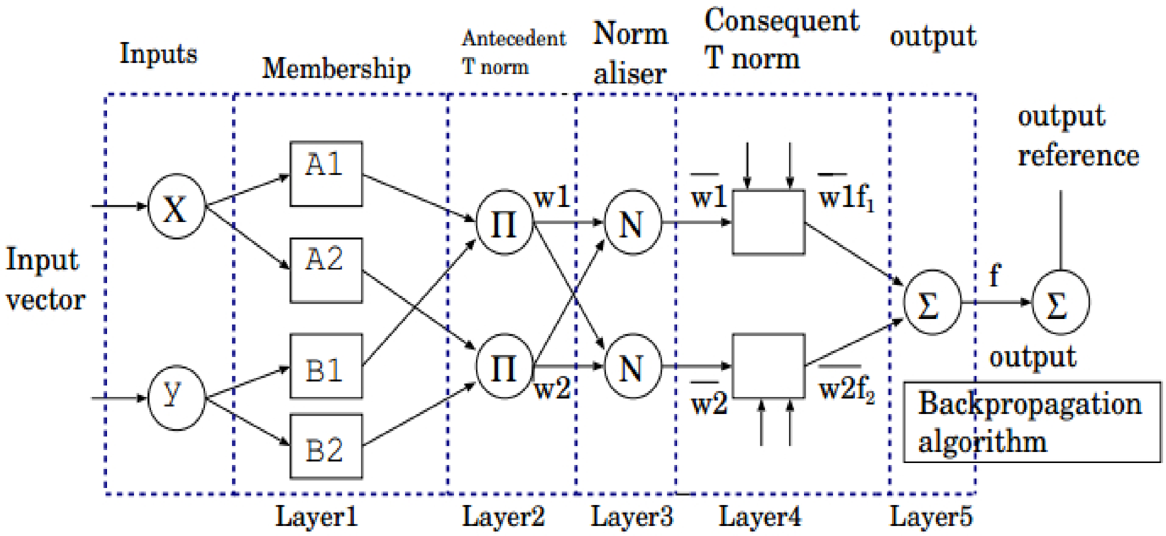

Figure 3: ANFIS architecture

The KNEA (Kernel-based nonlinear extension of the Arps decline model) model is a marvelous nonlinear model presented by [21] according to the Arps decline model and kernel method. In contrast to the non-parametric and “Black-Box” kernel-based models such as least-squares SVM, the KNEA model supports the “Grey-Box” idea and utilizes the semi-parametric approach to creating the nonlinear models [22]. The kernel-based methods indicate more efficiency when small samples are available [23, 24], while the KNEA model needs larger samples since samples are not accumulated in the model.

The MARS (Multivariate adaptive regression spline) model is a non-parametric regression method suggested by [25], which has no assumption on the relationships between the independent and dependent attributes. The modelling in this approach is based on some coefficients and functions. The basic function of the MARS model is the outcome of a truncated spline function or multiple spline functions. The nature of data specifies the number and characteristics of basic functions automatically. The MARS model is as good as the recursive auto-fractional regression method in dividing spatial regions, projection tracking method in processing high-dimensional data, and has the benefits of accumulative regression node self-adaptation.

Due to the better performance of ANFIS, SVM, KNEA, and MARS in prediction, they were used to perform the study, and then their results were considered to evaluate the performance.

3. EVALUATION CRITERIA

Accuracy of the regression equations for the derivation of PTFs was evaluated using R^2 (coefficient of determination) and RMSE (root mean square error) between the measured and predicted values and expressed in Eq. 1 and 2. In that equation, Y_i, X_i, and n are observed, predicated, and the number of samples values, respectively, where $Y(p_i)$ and $Y(o_i)$ are measured and predicted soil CEC values respectively; $\overline{Y_p}$ and $\overline{Y_o}$ are the means of measured and predicted soil CEC values, and n is the total number of observations.

$$R^2 = \left\{ \frac{\sum_{i=1}^{n}\left(Y(p_i) - \overline{Y_p}\right)\left(Y(o_i) - \overline{Y_o}\right)}{\sqrt{\sum_{i=1}^{n}\left(Y(p_i) - \overline{Y_p}\right)^2 \left(Y(o_i) - \overline{Y_o}\right)^2}} \right\} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[Y(pi) - Y(oi)]^2} \qquad (2)$$

RMSE values approaching zero and R values approaching 1 indicates that the model provides accurate predictions. R^2 is a statistic that will give some information about the goodness of fit of a model. In regression, the R^2 coefficient of determination is a statistical measure of how well the regression predictions approximate the real data points. When R^2 is equal to 1, it indicates that the regression predictions perfectly fit the data. The coefficient of the RMSE may be a measure of accuracy and reliability for calibration and test data sets [26]. Approximately 70% of the entire records used for training, therefore the testing sub-dataset included the remaining 30% of the records. All inputs and outputs were normalized. After normalization, data have the same order of magnitude. Without this step, and just in case of very different orders of magnitude between variables, small ones may have artificially lower influence during the training.

4. RESULTS AND DISCUSSION

Amini et al. [10] found that the neural network-based models provided more reliable predictions than the regression-based PTFs. The performance of different methods that are used by other studies to predict CEC represents briefly in table 3, according to the number of features they apply as input.

Table 3: The performance of different methods to predict CEC

References	Model	Input	R^2	RMSE
Amini et al. [10]	ANN	2	0.62	3.06
	MLR		0.54	3.50
Akbarzadeh et al. [12]	ANN	5	0.89	1.70
	MLR		0.72	5.32
	Neuro-fuzzy		0.97	0.87
Keshavarzi and Sarmadian [13]	ANN	2	0.94	0.47
	MLR		0.88	0.65
Shekofteh et al [14]	SVR	7	0.84	3.200
	ANFIS		0.81	3.380
	ANFIS	5	0.91	2.09

In this study, three features of pH, clay, and OM are applied as input to predict CEC. Statistical values of the various machine learning models with these input parameters during training and testing at the mentioned data are reported in table 4.

Table 4: Statistical values of the different machine learning models of study

model	Training		Testing	
	R^2	RMSE	R^2	RMSE
ANFIS	0.98	0.002	0.79	3.12
SVM	0.99	0.000	0.81	2.97
KNEA	0.96	0.01	0.80	3.003
MARS	0.99	0.009	0.81	3.000

However, the SVM, KNEA, and MARS models produced better CEC estimates compared with the other machine learning models of this study and others during the testing period. In all the methods, the RMSE of the testing phase was higher than RMSE during training, and no sign of overfitting is evident. This table suggested that the SVM and MARS models were the most stable models with the consistently smallest growth in RMSE during testing in contrast to training. The KNEA and ANFIS models also demonstrated an acceptable increase in testing RMSE. These results are consistent with other studies [14, 15] and show that new approaches can be applied to predict CEC efficiently and optimally. The best results of prior studies were gained with more inputs than the features of this study. It means that with dimension reduction and simplifying the problem, this method can obtain equivalent or better results.

5. ACKNOWLEDGMENTS

The authors would like to thank their friends, who provided the necessary data to do the study and left it at their disposal.

REFERENCES

[1] Rawlinson K. Microsoft's Bill Gates insists AI is a threat//BBC News. - 2015.
[2] Cellan-Jones R. Stephen Hawking warns artificial intelligence could end mankind//BBC news. – 2014.
[3] Bean R. How Big Data and AI are driving business innovation in 2018//MIT Sloan Management Review Webinar. - 2018.
[4] MILLER S. M. AI: Augmentation, more so than automation. - 2018.
[5] Wilson H. J., Daugherty P. R. Collaborative intelligence: humans and AI are joining forces//Harvard Business Review. - 2018. – Vol.96. – No.4. - pp.114-123.
[6] Agrawal A., Gans J., Goldfarb A. Prediction machines: the simple economics of artificial intelligence//Harvard Business Press. - 2018.
[7] Cearley D., et al. The top 10 strategic technology trends for 2018//Gartner. - 2018.
[8] Seybold C., Grossman R. B., Reinsch T. G. Predicting cation exchange capacity for soil survey using linear models//Soil Science Society of America Journal. - 2005. – Vol.69. _ No.3. - pp.856-863.
[9] Tamari S., Wösten J. H. M., Ruiz-Suarez J. C. Testing an artificial neural network for predicting soil hydraulic conductivity//Soil Science Society of America Journal. - 1996. – Vol.60. – No.6. - pp.1732-1741.
[10] Amini M., Abbaspour K. C., Khademi H., Fathianpour N., Afyuni M., Schulin R. Neural network models to predict cation exchange capacity in arid regions of Iran//European Journal of Soil Science. - 2005. – Vol.56. – No.4. - pp.551-559.
[11] Schaap M. G., Leij F. J. Using neural networks to predict soil water retention and soil hydraulic conductivity//Soil and Tillage Research. - 1998. – Vol.47. – No.1-2. - pp.37-42.
[12] Akbarzadeh A., Taghizadeh Mehrjardi R., Rahimi Lake H., Ramezanpour H. Application of artificial intelligence in modeling of soil properties// Environmental Research Journal. – 2009. – Vol.3. – No.2. – pp.19–24.
[13] Keshavarzi A., Sarmadian F. Comparison of artificial neural network and multivariate regression methods in prediction of soil cation exchange capacity (Case study: Ziaran region)//Desert. – 2010. – Vol.15. – No.2. – pp.167–174.
[14] Shekofteh H., Ramazani F. Prediction of soil cation exchange capacity using support vector regression optimized by genetic algorithm and adaptive network-based fuzzy inference system//Desert. – 2017. – Vol.22. – No.2. – pp.187-196.
[15] Shekofteh H., Ramazani F., Shirani H. Optimal feature selection for predicting soil CEC: Comparing the hybrid of ant colony organization algorithm and adaptive network-based fuzzy system with multiple linear regression//Geoderma. – 2017. – Vol.298. – pp.27–34.

[16] Specht D. F. A general regression neural network// IEEE Transactions on Neural Networks. – 1991 – Vol.2. – No.6. – pp.568–576.
[17] Jang J. S. R. ANFIS: adaptive-network-based fuzzy inference system// IEEE Transactions on Systems, Man, and Cybernetics. – 1993. – Vol.23. – No.3. – pp.665–685.
[18] Vapnik V. N. The nature of statistical learning theory// Statistics for Engineering and Information Science. - 2003.
[19] Yin Z., Wen X., Feng Q., He Z., Zou S., Yang L. Integrating genetic algorithm and support vector machine for modeling daily reference evapotranspiration in a semi-arid mountain area//Hydrology Research. – 2017. – Vol.48. – No.5. – pp.1177–1191.
[20] Mehdizadeh S., Behmanesh J., Khalili K. Using MARS, SVM, GEP and empirical equations for estimation of monthly mean reference evapotranspiration// Computers and Electronics in Agriculture. – 2017. – Vol.139. – pp.103–114.
[21] Ma X., Liu Z. Predicting the oil production using the novel multivariate nonlinear model based on Arps decline model and kernel method// Neural Computing and Applications. – 2018. – Vol.29. – No.2. – pp.579–591.
[22] Ma X., Xie M., Wu W., Zeng B., Wang Y., Wu X. The novel fractional discrete multivariate grey system model and its applications// Applied Mathematical Modelling. – 2019. – Vol.70. – pp.402–424.
[23] Ma X., Liu Z. The GMC (1, n) model with optimized parameters and its application// Journal of Grey System. – 2017. – Vol.29. – No.4. – pp.122–138.
[24] Ma X. A brief introduction to the Grey Machine Learning//Journal of Grey System. – 2019. – Vol.31. – No.1. - pp.1-12.
[25] Friedman J. H. Multivariate adaptive regression splines//The Annals of Statistics. – 1991. – Vol.19. – No.1. – pp.1–67.
[26] Ozturk M., Salman O., Koc M. Artificial neural network model for estimating the soil temperature//Canadian journal of soil science. - 2011. – Vol.91. – No.4. – pp.551-562.