POLISH JOURNAL OF SOIL SCIENCE
VOL. LI/2 2018 PL ISSN 0079-2985

DOI: 10.17951/pjss/2018.51.2.185

PEYMAN AMIN*, RUHOLLAH TAGHIZADEH-MEHRJARDI**,
ALI AKBARZADEH***, MOSTAFA SHIRMARDI**

COMPARISON OF DATA MINING TECHNIQUES TO PREDICT AND MAP THE ATTERBERG LIMITS IN CENTRAL PLATEAU OF IRAN

Received: 22.03.2018
Accepted: 12.06.2018

Abstract. The Atterberg limits display soil mechanical behavior and, therefore, can be so important for topics related to soil management. The aim of the research was to investigate the spatial variability of the Atterberg limits using three most common digital soil-mapping techniques, the pool of easy-to-obtain environmental variables and 85 soil samples in central Iran. The results showed that the maximum amount of liquid limit (LL) and plastic limit (PL) were obtained in the central, eastern and southeastern parts of the study area where the soil textural classes were loam and clay loam. The minimum amount of LL and PL were related to the northwestern parts of the study area, adjacent to the mountain regions, where the samples had high levels of sand content (>80%). The ranges of plasticity index (PI) in the study area were obtained between 0.01 to 4%. According to the leave-in-out cross-validation method, it should be highlighted the combination of artificial bee colony algorithm (ABC) and artificial neural network (ANN) techniques were the best model to predict the Atterberg limits in the study area, compared to the support vector machine and regression tree model. For instance, ABC-ANN could predict PI with RMSE, R^2 and ME of 0.23, 0.91 and -0.03, respectively. Our finding generally indicated that the proposed method can explain the most of variations of the Atterberg limits in the study area, and it could be

* Faculty of Natural Resources and Desert Studies, Yazd University, Iran; corresponding author: peymanamin50@yahoo.com

** Faculty of Agriculture and Natural Resources, Ardakan University, Iran.

*** Faculty of Agriculture, Shahrekord University, Iran.

recommended, therefore, as an indirect approach to assess soil mechanical properties in the arid regions, where the soil survey/sampling is difficult to undertake.

Keywords: Atterberg limits, artificial bee colony, artificial neural networks, support vector machine, regression tree

INTRODUCTION

The Atterberg limits are a criterion for explanation of soil moisture content. Based on this criterion, three limits are defined for moisture in the soil, including shrinkage limit (SL), plastic limit (PL), and liquid limit (LL) (Atterberg 1911). Fine-grained soils can show different status according to the amount of their water absorption. These three boundaries are among the four states of soil behavior, which are solid, semi solid, plastic, and liquid. The Atterberg limits are mainly used to name and classify soils (Campbell 2001), to quantify the physical activity of soils for engineering purposes and to estimate the other soil mechanical parameters such as shear strength, bearing capacity, compressibility, swelling potential and specific surface area (De Jong *et al*. 1990, Fanourakis 2012, Moradi 2013, Saikia *et al*. 2017).

Atterberg moisture limits (SL, PL and LL) are among the most important characteristics of soils which provide soil researchers with valuable information about soil behavior, especially in soil management discussions (Keller and Dexter 2012). However, little attention appears to have been paid to their prediction using easily measurable soil properties. For example, Zolfaghari *et al*. (2015) predicted the Atterberg limits by some soil characteristics and bio-environmental data using an ANN method. Their results showed that the soil organic matter, clay content, calcium carbonate equivalent, and terrain attributes could explain most of the variances of the Atterberg limits in the region. Keller and Dexter (2012) similarly reported that the PL had significant correlation with the soils having more than 35% clay content in agricultural soils in different countries. Mirkhani *et al.* (2006) showed that LL and PL had significant correlation with cation exchange capacity (CEC), bulk density and water saturation percentage. Tol *et al.* (2016) found a positive correlation between clay content and the Atterberg limits, a negative correlation between sand content and the PI, and no significant correlation between organic carbon and PI in soils of South Africa. However, Zentar *et al*. (2009) indicated the significant effect of soil organic carbon on the Atterberg limits. The similar positive correlation between soil organic carbon and PI was reported by Abdi *et al*. (2018), who investigated the Atterberg limits in forest soils of Hyrcanian forest in northern Iran.

Digital soil mapping represents a set of computer calculations for predicting the distribution of soils in different landforms. This technique has started its growth from the earliest days of soil survey studies and has evolved along with

the advances made in the processing of information (Scull *et al.* 2003). Digital soil mapping compiles and creates systems for soil spatial information, which can assist users in deciding to address environmental and agricultural issues (Lagacherie and McBratney 2007). One of the main aspects of digital soil mapping is the use of various models (i.e. data mining techniques) to simplify the complexity of the soil system. Therefore, soil-landform models represent a simplified form of complex relationships between soil and landform which depicts the evolutionary processes of the soil and its pattern of distribution (Grunwald 2006), also in order to deal with the global issues including climate change, land degradation, biodiversity loss, etc., detailed accurate spatial soil information is urgently needed (Gan-lin *et al.* 2017).

Data mining is a method of practical discovery of a meaningful pattern, form and process in a manner of sifting from data using different pattern recognition techniques. Sifting is a method by which records are moved to allow other records to be entered (Mena 1999). Several broad types of data mining models have been used for digital mapping of soil classes and properties (Brungard *et al.* 2015), such as logistic regression (Hengl *et al*. 2007, Jafari *et al*. 2012, Marchetti *et al*. 2011, Zeraatpisheh *et al*. 2017), classification trees (Bui and Moran 2003; Kim *et al*. 2012), random forests (Stum *et al.* 2010, Poggio *et al*. 2013, Pahlavan Rad *et al.* 2014, Zeraatpisheh *et al*. 2017), neural networks (Behrens *et al.* 2005, Moonjun *et al*. 2010; Jafari *et al.* 2013, Taghizadeh-Mehrjardi *et al*. 2016), and support vector machines (Kovačević *et al.* 2010).

The artificial bee colony (ABC) algorithm which is one of the data mining techniques was first introduced by Karaboga in 2005 to optimize mathematical functions. In this way, each answer, which is one place in search space, represents a potential food area and the quality of the answer is equivalent to the quality of the food source. The agents (artificial bees) are looking for and exploiting food resources in the search space (Panigrahi *et al.* 2011). ABC method uses three types of agents including employed bees (EB), onlooker bees (OB) and scout bees (SB). EB are related to the current algorithm's answers. At each step of the algorithm, the EB tries to improve the answer that it delivers through a local search step. EB will then attempt to use the OB for its current location. OB chooses improved locations according to their quality. This means that better answers will attract more OB. If the OB that was assigned to work was able to find a better place, the EB will update its location; otherwise, the EB will remain in its current location. Additionally, the EB will leave its location if it is not able to improve its position in a certain number of steps, which is called *limit*. If the EB abandons its location, it will become a discoverer bee. This means that the EB chooses a new place randomly in the search space (Panigrahi *et al*. 2011).

There are a large number of available DSM techniques (Kuhn and Johnson 2013), but a little attempt has been made to combine data mining techniques and feature selection algorithms in DSM (Taghizadeh-Mehrjardi *et al*. 2017).

Feature selection algorithms condense the feature spaces and, hence, they might indirectly have resulted in increasing prediction accuracies. To the best of our knowledge, there is no published paper about application of ABC and ANN techniques to predict and map the Atterberg limits in arid regions. Therefore, the main objective of this research was to explore whether prediction improvements could be achieved using a feature selection technique (artificial bee colony algorithm) compared with artificial neural network, support vector machine and regression tree models fed with all auxiliary variables. Moreover, we tried to answer which auxiliary variables are the most important parameters to explain the spatial variations of the Atterberg limits.

MATERIALS AND METHODS

Description of the study area

The study area which is a part of Yazd-Ardakan region or plain has a total area of 15,950 km^2 and geographic coordinates are between 31°48' and 32°13'N and 52°57' to 54°59'E. This region is located in the northern part of the Yazd province in Central Plateau of Iran (Fig. 1). The study area includes about 24.9% of the total area of the Yazd province. This rectangular area is surrounded by the various mountain ranges including Shirkooh, Ahangaran, Morg-e-Zard, Haft Adamin, Koonza, and Chek Chek. The area is extended to the Siahkooh hole with a common southeast-northwest slope (Ekhtesasi *et al.* 1995).

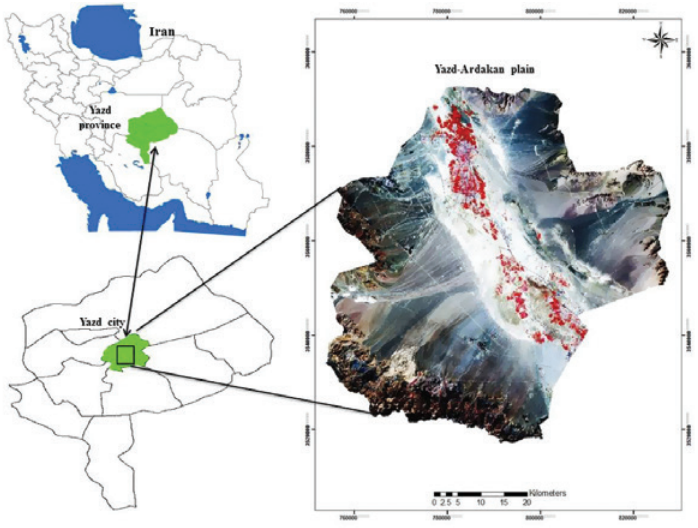

Fig. 1. Location of the study area in the Yazd-Ardakan region

Soil sampling and laboratory analysis

The 85 samples were taken from soil surface (0–20 cm) in three different geomorphological units including coarse, medium and fine-grained plains. The configuration of the sampling locations was based on the conditioned Latin hypercube method (Minasny and McBratney 2006) using a number of environmental variables that were found to have the most variation within the area, including geomorphological units, elevation, wetness index, slope, three first spectral bands of Landsat image, and normalized difference vegetation index. Latin hypercube method is a stratified random procedure that provides an efficient way of sampling variables from their multivariate distributions. This methodology is used in soil and environmental studies to evaluate uncertainty in the prediction models. Due to high sand contents (more than 90%) in the 45 collected soil samples, we could only measure the Atterberg limits in 40 out of 85 soil samples. Consequently, we run the spatial models with 40 data. Figure 2 further illustrates the sampling points over the study area. Most of the sampling points were located in the middle, eastern and southeastern parts of the study area (Fig. 2). These regions were mostly plains – agricultural lands and lands covered by scattered plants (*Haloxylon*). The remaining parts of the study area included bare plain and plain appendage such as barren lands containing dry rivers, desert pavements and rangelands.

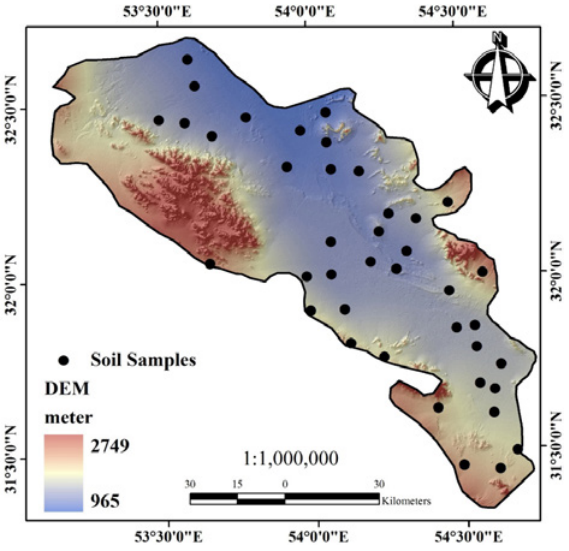

Fig. 2. Position of the sampling points over the study area

The liquid limit (LL) and plastic limit (PL) were determined by Casagrande and Wick methods, respectively (British Standard 1975). The plasticity index (PI) was then calculated from the difference between these two limits (Equation 1):

$$PI = LL - PL \quad (1)$$

Soil textural classes and soil particle size distribution were determined by the dry sieving method. Particles with size smaller than two millimeters were calculated by the hydrometer method (Gee and Bauder 1986).

Auxiliary data

Two sources of information, which are mainly used in digital soil mapping, were applied in the present study (Table 1). These two sources of information were Landsat satellite images and digital elevation model (DEM). An ASTER digital model with a spatial resolution of 30 meters was used in this research. After preparing a DEM, it was used to extract the auxiliary data or estimated images of soil genetic variables. Different parameters of landscape, including slope (SLOP), elevation (ELV), height above drainage network (HADN), modified catchment area (MCA), middle slope position (MSP), valley depth (VD), wetness index (WI), multi-resolution valley bottom flatness index (MrVBF), and watershed slope (WSLOP) were determined and extracted in the environment of System for Automated Geoscientific Analyses (SAGA) software which is a Geographic Information System (GIS) computer program. Extraction of all of the mentioned parameters was carried out based on method introduced by Hengl *et al.* (2004).

Different types of soils have different spectral properties (Andronikov and Dorbrolvskiy 1991). In this study, changes in the characteristics and type of soils could be detected easily by satellite data because most of the study area did not have vegetation cover (Metternicht and Zinck 2003). The data derived from ETM^+ sensor of Landsat satellite was used to study some soil formation factors. Based on soil formation factors detected in the study area, some indices such as normalized difference vegetation index (NDVI), clay index (CI), carbonate index (CaI), gypsum index (GI), salinity index (SI), and brightness index (BI) were determined.

Table 1. Land surface parameters used for spatial prediction of Atterberg limits

Auxiliary data	Land surface parameters	Definition
Terrain attributes	Slope (SLOP)	Average gradient above flow path
	Elevation (ELV)	Height above sea level (m)
	Height above drainage network (HADN)	Relative height above depth
	Modified catchment area (MCA)	Calculates the flow accumulation
	Mid-slope position (MSP)	Calculates the extent that each point
	Valley depth (VD)	Metres
	Wetness index (WI)	Ln (FA/SG)
	Multi-resolution valley bottom flatness index (MrVBF)	Measure of flatness and upness

Auxiliary data	Land surface parameters	Definition
	Watershed slope (WSLOP)	Average gradient above flow path in watershed
Remote sensing data	Six ETM bands	B1, B2, B3, B4, B5, B7
	Normalized difference vegetation index (NDVI)	(B4-B3)/(B4+B3)
	Clay index (CI)	B5/B7
	Carbonate index (CaI)	B3/B2
	Gypsum index (GI)	(B5-B4)/(B5+B4)
	Salinity index (SI)	(B3-B4)/(B2+B4)
	Brightness index (BI)	$((B3)^2+(B4)^2)^{0.5}$

Feature selection using the artificial bee colony (ABC) algorithm

The process of selecting an attribute is usually used for issues where data includes many features. This process leads to a reduction in the vector dimensions of the studied properties. It is applicable by removing unnecessary features and selecting the essential features to learn the model. This process ultimately improves predictive accuracy and improves the predictive capabilities of predictive models. The process of choosing a feature is, in fact, a matter of selecting a subset of features that is sufficient and necessary to explain the intended purpose (Oreski *et al.* 2012). Here, we used the ABC algorithm to rank the most important predictors. The main steps in the ABC algorithm are as follows (Karaboga and Bahriye 2009): (1) creation of a primary population, (2) repeat, (3) establishment of employed bees (EB) on their food sources, (4) establishment of onlooker bees (OB) on food sources in terms of their nectar, (5) sending the scout bees (SB) to the search space to find new food sources, (6) remembering the best source of food found up until that time, and (7) establishing a stopping condition. The ABC analysis was implemented in the WEKA software (Hall *et al.* 2009).

In the first step, ABC produces an initial population randomly. It means that SN produces an answer or response. SN denotes the number of EB or OB. Every answer is x_i (i = 1, 2,…, SN). A D-dimensional vector, where *D* is the number of optimization parameters. After this stage, the population of positions (responses or answers) is exposed to repeated cycles (c = 1, 2, …, MCN) of search processes done by EB, OB, and SB. An artificial OB selects the source of food depending on the probability (P_i) associated with that source of food. P_i can be estimated by the following formula (Equation 2):

$$P_i = \frac{fit_i}{\sum_{N=1}^{SN} itn} \tag{2}$$

Where, fit_i is the fitness level of each response of i, which is a proportion of the volume of nectar in the food source at i. Also, SN is the number of food sources, which is equal to the number of EB or OB. ABC uses the following formula (Equation 3) in order to create a position of food for the candidate from the place of the old food:

$$v_{ij} = x_{ij} + r_{ij}\left(x_{ij} - x_{kj}\right) \tag{3}$$

Where, $k \in \{1, 2, \ldots, SN\}$ and $j \in \{1, 2, \ldots, D\}$ are randomly selected. Although k is randomly determined, it must be different with i. Also, r is a random number between [-1, 1]. If the value of the generated parameter in this way exceeds its predetermined value, then the parameter can take an acceptable value. For example, it can take the limit value (Karaboga and Bahriye 2009).

The reason for choosing the ABC algorithm in the present study was its simplicity, reputation and its good application among various bee algorithms. It is necessary to note that the efficiency of an algorithm depends on the type and structure of the problem. Therefore, a very successful algorithm in a particular problem can have a poor performance in another problem. Consequently, the field of Meta-Algorithms is highly experimental, and it is not possible to determine the precision of a specific algorithm prior to the experimental experiments to solve a particular problem.

Spatial prediction models

Artificial neural networks (ANNs)

An artificial neuron similar to a biological nerve consists of three basic parts, including the input, the core and the output. A matrix of numbers is defined as the input of the network in an artificial neuron, which plays the same role as inputs in a biological neuron. Inside the nucleus of a biological neuron, different chemical processes are carried out. The equivalent of these processes in an artificial neuron is a function called *the activation function*, which is represented by *f*. Communications between neurons in an artificial network are also defined by adjusting the weights for each neuron. The output of the artificial neuron is also the result of applying the function *f* to the linear combination of inputs (Dixon and Candade 2008). In the train step of ANNs, the optimum number of neurons in the hidden layer was fixed based on multiple runs in a trial-and-error strategy and the root-mean-square error (RMSE) as the criterion. We also use the Levenberg–Marquardt training algorithm (Levenberg 1944, Marquardt 1963) due to its efficiency and simplicity.

Regression tree (RT)

Tree classification and regression model is a non-parametric algorithmic method which is able to predict quantitative variables or classified variables based on a set of quantitative and qualitative predictor variables. In this method, a set of logical conditions is used as a tree structure algorithm for quantization or quantitative prediction of a variable. Creating a decision tree has two steps. The first step is the stage of tree creation and growth which includes the bonding and splitting. The second step is the stop and pruning stage. The purpose of the second step is to minimize the estimated error. In the present study, cubist software (Quinlan 2001) was used to construct the decision tree for predicting the Atterberg limits of soil samples.

Support vector machine (SVM)

Support vector machine (SVM) is based on the theory of statistical learning, which was first used in the 1960s. This method is a supervised non-parametric technique (Pao 1989). In this method, specimens that form the boundaries of classes are obtained using all the bands and an optimization algorithm, and using them, an optimal linear decision boundary for separating classes is computed. These specimens are called *support vectors* (Keshavarz *et al.* 2004). For this research, the radial basis function (RBF) kernel, as suggested by Hsu *et al.* (2009), was used as the kernel functions to act as a universal function approximator. We also used simulated annealing algorithm for optimizing the parameters of SVR.

Model evaluation

A validation practice was done by deleting the data and predicting it to evaluate the efficiency of the model in order to predict the Atterberg limits of soil samples. In this method, division is performed repeatedly unlike the usual methods available for data splitting. This operation improves the efficiency of this method. In such circumstances, the best option for validating the model of digital mapping is using a data deletion and its prediction technique. The database set (n) is divided into n-1 position for calibration and one position for validation. In each replication, the model runs for the deleted position and the deleted variable is predicted. Then it is compared with the real value and the estimated error can be calculated. This process is performed for all sampling locations. Correlation coefficient for real and estimated values of Atterberg limits were determined after designing a suitable model. Also, each model was validated with calculation the root-mean-square error (RMSE) and the mean error (ME).

RESULTS AND DISCUSSION

Statistical summary of measured data

Table 2 shows some statistical characteristics for percentage of clay, silt and sand in 85 samples taken from the study area. According to these results, the amount of sand in the study area was very high. Thus, only 40 soil samples were finally selected to determine the Atterberg limits. The locations of these 40 points are shown in Figure 2. The textural classes of 40 soil samples used for determination of the Atterberg limits are presented in Table 3. The soil textural classes over the study area were sandy clay loam, loam, clay loam, and sandy loam according to the soil texture triangle. The maximum and minimum numbers of soil textural classes were related to sandy clay loam and clay loam, respectively. Other remaining soil samples (45 samples) that were not tested for the Atterberg limits had sandy and loamy sand textural class. The percentage of sand in these samples was more than 80%. Hence, these soil samples did not come in the form of mud by absorbing water. For a soil to be within the Atterberg limits, its clay and silt content must be high. In general, soils with clay content below 10% do not reach the Atterberg limits (Keller and Dexter 2012).

Table 2. Some statistical characteristics for clay, silt and sand in 85 samples

Statistic	**Clay (%)**	**Silt (%)**	**Sand (%)**
Minimum	1	2	29
Maximum	41	50	95
Mean	16.24	14.91	68.84
Standard deviation	9.38	10.36	16.33

Table 3. The number of soil textural classes in the study area

Soil Textural Classes	Number
Sandy clay loam	14
Loam	7
Clay loam	6
Sandy loam	13

Feature selection using the artificial bee colony (ABC) algorithm

The results of the ABC algorithm showed that some of the auxiliary variables were more important for entering the modeling process. These variables included NDVI, WI, MrVBF, SI, CI, and Band 7 (i.e. reflectance value of band 7: shortwave IR-2 with band width of 2.09–2.35 μm) derived from ETM^+ sensor of Landsat satellite. The results showed that the Band 7 of Landsat images were

of greatest importance for entering the modeling process by the ABC algorithm. The reason for the significant importance of the band 7 of Landsat images was probably due to its high correlation with soil moisture content. Soil moisture content has been shown to correlate directly with the soil texture and, hence, the amount of the Atterberg limits (Liao *et al.* 2013). In addition, various researchers have confirmed the effectiveness of satellite imagery to estimate the clay content and, consequently, the Atterberg limits (Alavi-Panah *et al.* 2008, Liao *et al.* 2013, Ahmed and Iqbal 2014).

Wetness index (WI) is a suitable indicator used to identify areas susceptible to moisture storage in a region. Therefore, by this indicator, it is possible to identify areas with high clay content. Consequently, WI helps to determine the amount of the Atterberg limits in a region. Figure 3 presents the spatial variation of WI over the study area. According to this figure, most of the area in the region due to the presence of abundant dry and semi-dry water ways, had maintained high moisture content on its soils and, finally, the value of WI increased in the study area.

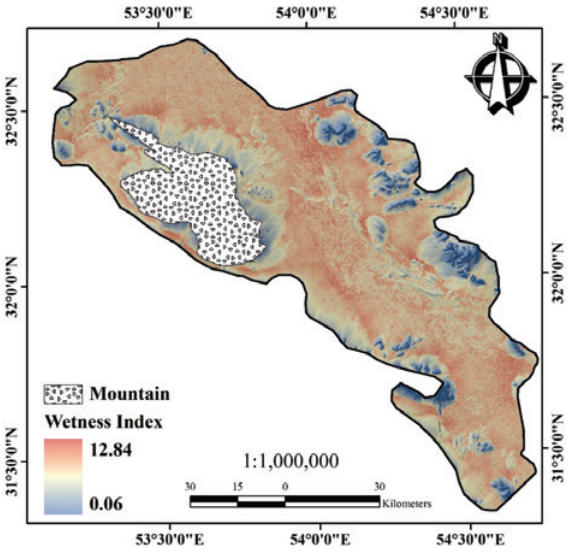

Fig. 3. Spatial variation of wetness index (WI) over the study area

Figure 4 shows the spatial variation of normalized difference vegetation index (NDVI) over the study area. As shown in this figure, most of the vegetation cover in the study area belonged to the central parts. These areas, which include thin vegetation cover consisting of *Haloxylon*, agricultural lands and fruit gardens, are marked in red. Most of the area in the eastern, western and northern parts of the region, due to the coarse soil textural class, medium to high pavement cover, and adjacent parts of the mountains, were barren lands. In some parts, there

was also small and scattered vegetation, especially as a form of rangeland. It is worth noting that there was a significant correlation between NDVI and LL and PL, and, therefore, it could help indirectly in the prediction of the Atterberg limits (Table 4). Table 4 summarized the correlation of auxiliary variables and the Atterberg limits. This further indicated that there is a significant correlation between the Atterberg limits and some other predictor variables (e.g. WI, MrVBF, HADN, SI, CI, and the six spectral bands derived from ETM$^+$ sensor of Landsat satellite). These results somehow confirmed the finding obtained by the ABC.

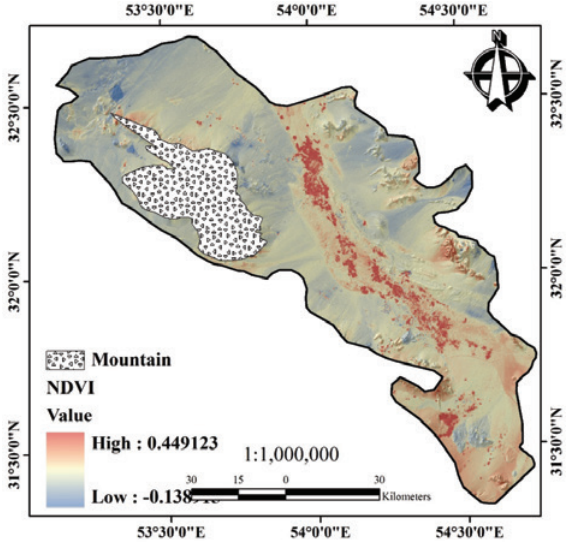

Fig. 4. Spatial variation of normalized difference vegetation index (NDVI) over the study area

Table 4. Correlation coefficients between the Atterberg limits and auxiliary variables (*, **: meaning in the level of 95% and 99%, respectively)

	B1	B2	B3	B4	B5	B7	NDVI	CI	CaI	GI	SI
LL	0.28*	0.25*	0.29**	0.34**	0.37**	0.45**	0.50**	0.47**	0.21*	0.18	0.41**
PL	0.25*	0.18	0.11	0.30**	0.34**	0.48**	0.47**	0.44**	0.14	0.11	0.31**

	BI	SLOP	ELV	HADN	MCA	MSP	VD	WI	MrVBF	WSLOP
LL	0.14	-0.46**	0.49**	0.11	0.16	-0.35**	0.07	0.52**	0.38**	-0.25*
PL	0.16	-0.42**	0.52**	0.14	0.12	-0.33**	0.11	0.55**	0.33**	-0.28*

Spatial prediction models

Different types of data mining models including ANN, RT and SVM were accomplished using two input categories. These two input categories were all input data and data selected by the ABC algorithm. These results are presented in Tables 5 and 6.

Table 5. Statistical properties of modelling with all inputs data

Atterberg Limits	Model	R^2	RMSE	ME
Liquid limit (LL)	ANN	0.83	1.73	-0.11
	SVN	0.79	1.75	-0.12
	RT	0.80	1.74	-0.11
Plastic limit (PL)	ANN	0.80	1.70	-0.09
	SVN	0.75	1.83	-0.12
	RT	0.76	1.76	-0.09
Plasticity index (PI)	ANN	0.89	0.25	-0.05
	SVN	0.85	0.30	-0.06
	RT	0.85	0.28	-0.04

Table 6. Statistical properties of modelling with selected data by the ABC algorithm

Atterberg Limits	Model	R^2	RMSE	ME
Liquid limit (LL)	ANN	0.85	1.62	-0.09
	SVN	0.80	1.72	-0.11
	RT	0.81	1.70	-0.10
Plastic limit (PL)	ANN	0.85	1.56	-0.07
	SVN	0.78	1.80	-0.10
	RT	0.79	1.75	-0.08
Plasticity index (PI)	ANN	0.91	0.23	-0.03
	SVN	0.87	0.27	-0.04
	RT	0.87	0.24	-0.03

The results indicated that using the feature selection algorithm improved prediction of all parameters related to the Atterberg limits (LL, PL, and PI) in all three methods (ANN, SVM, and RT) (Tables 5 and 6). For example, the ANN improved the accuracy of prediction for LL, PL and Pl by 11%, 4% and 2%, respectively. Similarly, the feature selection algorithm improved the performance of SVM and RT models by 3% for the Atterberg limits. Generally, using the feature selection algorithm simplified the model and rendered the model less time-consuming. In addition, the ABC algorithm was able to properly select the most useful variables. Moreover, this model (ABC algorithm) was able to remove the additional data from the modeling process and ultimately led to a simpler model. Depending on the RMSE values, it was found that the ABC algorithm and ANN models were the best models for predicting the Atterberg limits. Therefore, the ABC algorithm model was used for digital mapping of the Atterberg limits with acceptable accuracy over the study area. Similarly, Zolfaghari *et al*. (2015) estimated the Atterberg limits by a data mining technique (ANN method) using a set of input data, including soil properties (clay, organic matter and equivalent calcium carbonate), topography indices (elevation, slope and curve map) and vegetation index (NDVI) in parts of western Iran.

Results showed that the ANN method was able to predict the Atterberg limits with a high accuracy. Sherzoy (2017) compared two models including SVM and ANFIS for predicting the Atterberg limits with 54 soil samples from the area of Peninsular Malaysia. The outcome of his study showed that the ANFIS model had higher accuracy than the SVM model. Mukhlisin and Rahman (2014) predicted the Atterberg limits via the ANN and ANFIS models with 54 soil samples across Peninsular Malaysia. The results showed that the ANFIS model produced more accurate results than the ANN model in terms of liquid limit and plasticity index, but not plastic limit. Yildirim and Gunaydin (2011), using multiple regression analysis and artificial neural networks, tried to predict the Atterberg limits in the public highways of Turkey's different regions. Results showed that regression analysis and artificial neural network estimation indicated strong correlations ($R^2 = 0.80–0.95$) between the sieve analyses, the Atterberg limits, maximum dry unit weight (MDD) and optimum moisture content (OMC).

As an additional visual analysis of the validation, Figure 5 shows the comparison of the measured values of a) liquid limit, b) plastic limit and c) plasticity index with the estimated values from composition of artificial bee colony and artificial neural network model which was selected as the best model. The dotted line represents the 1:1 line. If all the measured points and estimates of the model are equal, then the point is focused on the 1:1 line. Deviation from this line indicates the amount of error in the map. Therefore, with the conformity of the fitted line with the dotted line and strong determination coefficient of 0.85 to 0.86, the prepared maps of the Atterberg limits are highly accurate.

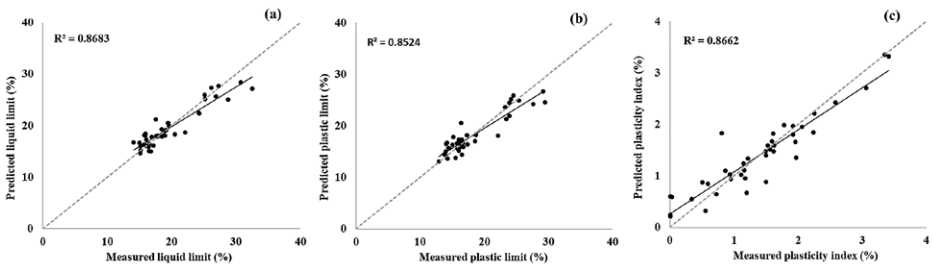

Fig. 5. Comparison of measured Atterberg limits with the values derived from the model: a) liquid limit, b) plastic limit and c) plasticity index

Spatial distribution of the Atterberg limits as a result of ABC-ANN

The spatial variation of liquid limit (LL) over the study area is illustrated in Figure 6. The variation range for the LL values in the study area was between 14% and 33%. As shown in Figure 6, the highest amount of LL changes was between 15% and 25%. The highest amount of LL values was found in the central, eastern and southeastern parts of the study area where the soil textural class

was loam and clay loam. These areas were often covered by vegetations and were clay plains with fine grained soils. The least amount of LL was related to the northwest (adjacent to the mountains), northeast, and south of the region. In these areas, the sand content and coarse-grained sediments is potentially high due to the large number of active and semi-active channels. The surface fine-grained sediments are washed and transferred to the middle parts of the region by most of these waterways.

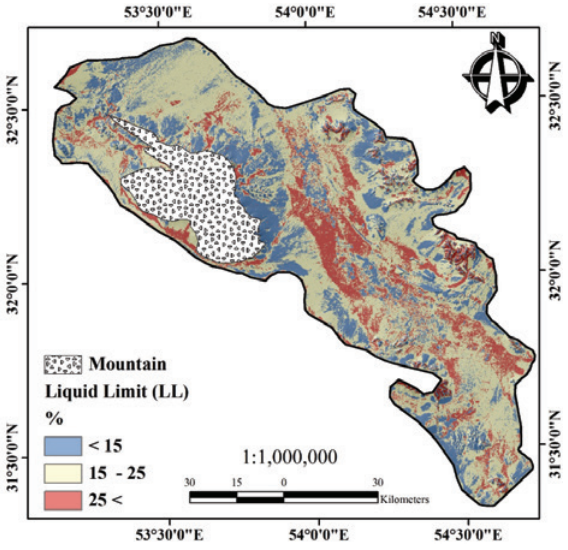

Fig. 6. Spatial variation of liquid limit (LL) over the study area

The spatial variation of plastic limit (PL) over the study is shown in Figure 7. The variation range for PL in the study area was between 13% and 30%. According to Figure 7, the highest amount of PL, ranging from 15% to 25% (similar to the LL), was distributed throughout the region. The highest amount of PL values was found in the central, eastern and southeastern parts of the study area similar to the LL. The least amount of PL was related to the northwest (adjacent to the mountains), northeast, and south of the study area, as well.

The spatial variation of plasticity index (PI) over the study is shown in Figure 8. The variation range for PI in the study area was between 0.01% and 4%. According to Figure 8, the highest amount of PI, ranging from 0.01% to 2.5%, was distributed throughout the region. In addition, in some parts of the study area, especially in northeastern and eastern parts of the region, this range was between 2.5% and 4%. The very little difference between LL and PL levels was due to the low clay content in the study area which could not increase the amount of PI. Even the highest amounts of clay content that was found in the central parts of the study area (ranging from 25% to 41%) could not increase

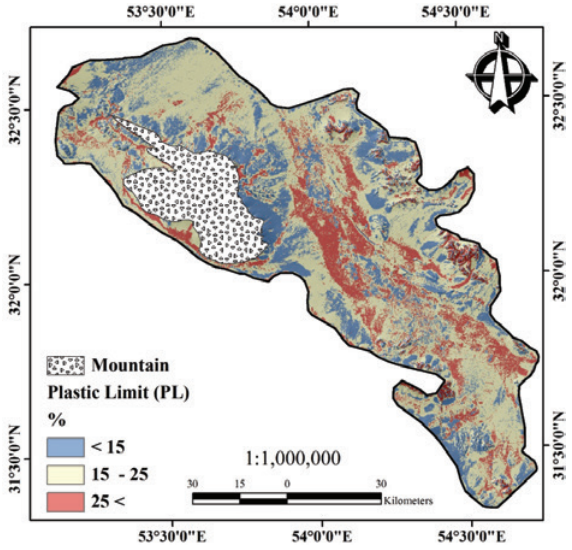

Fig. 7. Spatial variation of plastic limit (PL) over the study area

PI by more than 4% in the region. Rahimnia and Heidari Bani (2010) also got similar results by reviewing PI levels of some soil samples taken from Yazd and Meybod regions of central Iran. They also attributed the low amount of PI to low levels of clay content in the study area.

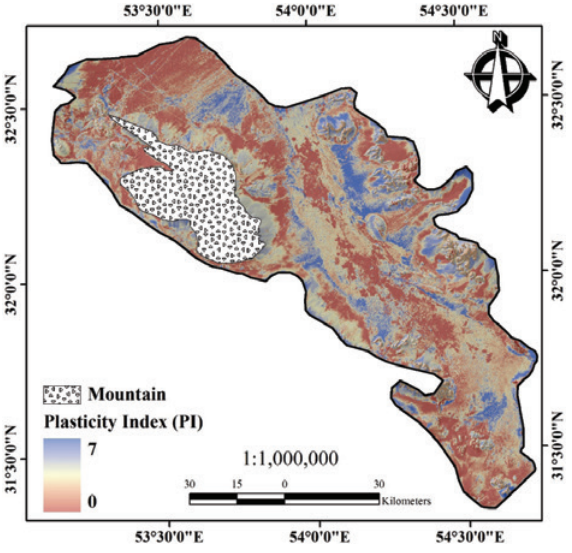

Fig. 8. Spatial variation of plasticity index (PI) over the study area

CONCLUSIONS

In our study, we tried to digitally predict and map the Atterberg limits using different data mining techniques in Yazd-Ardakan region, central Iran. Our results indicated that the combination of artificial bee colony and artificial neural network method was the best model to predict the Atterberg limits in the study area from all selected data mining techniques. The results also showed that the maximum amount of liquid limit and plastic limit were found in the central, eastern and south eastern parts of the study area where the soil textural classes were loam and clay loam. Overall, that application of soft computing techniques could be recommended to prepare spatial maps of soil properties using limited soil data particularly in arid regions where soil sampling is very difficult and time-consuming.

REFERENCES

[1] Abdi, E., Babapour, S., Majnounian, B., Zahedi-Amiri, G., Deljouei, A., 2018. *How does organic matter affect the physical and mechanical properties of forest soil?* Journal of Forest Research, 29(3): 657–662.

[2] Ahmed, Z., Iqbal, J., 2014. *Evaluation of Landsat TM5 multispectral data for automated mapping of surface soil texture and organic matter in GIS*. European Journal of Remote Sensing, 47: 557–573.

[3] Alavi-Panah, S.K., Goossens, R., Matinfar, H.R., Mohamadi, H., Ghadiri, M., Irannejad, H., 2008. *The efficiency of Landsat TM and ETM+ thermal data for extracting soil information in arid regions*. Agricultural Science and Technology, 10: 439–460.

[4] Andronikov, V.L., Dorbrolvskiy, G.V., 1991. *Theory and methods for the use of remote sensing in the study of soils*. Remote Sensing, 28: 92–101.

[5] Atterberg, A., 1911. *Physical soil examination about the plasticity of clays*. International Communication for Soil Science, 1: 10–43.

[6] Behrens, T., Forster, H., Scholten, T., Steinrucken, U., Spies, E.D., Goldschmitt, M., 2005. *Digital soil mapping using artificial neural networks*. Plant Nutrition and Soil Science, 168: 21–33.

[7] British Standard. 1975. *Methods of testing soils for civil engineering purposes*. British Standard Institute, London.

[8] Brungard, C.W., Boettinger, J.L., Duniway, M.C., Wills, S.A., Edwards Thomas, C.J., 2015. *Machine learning for predicting soil classes in three semi-arid landscapes*. Geoderma, 240: 68–83.

[9] Bui, E.N., Moran, C.J., 2003. *A strategy for filling gaps in soil survey over large spatial extents: An example from the Murray-Darling basin of Australia*. Geoderma, 111: 21–44.

[10] Campbell, D.J., 2001. Liquid and plastic limits. In: K.A. Smith, C.E. Mullins (eds.), *Soil and Environmental Analysis-Physical Methods*, Dekker Inc., New York, pp. 349–375.

[11] De Jong, E., Acton, D.F., Stonehouse, H.B., 1990. *Estimating the Atterberg limits of Southern Saskatchewan soils from texture and carbon contents*. Canadian Journal of Soil Science, 70: 543–554.

[12] Dixon, B., Candade, N., 2008. *Multispectral land-use classification using neural networks and support vector machines: One or the other, or both?* International Journal of Remote Sensing, 29(4): 1185–1206.

[13] Ekhtesasi, M.R., Ahmadi, H., Baghestani, N., Khalili, A., Feiznia, S., 1995. *Seeking the origin of sand dunes in Yazd-Ardakan plain* (in Persian). Research Institute of Forests and Rangelands, p. 250.
[14] Fanourakis, G.C., 2012. *Estimating soil plasticity properties from pedological data*. The South African Institution Journal of Civil Engineering, 2: 117–125.
[15] Gan-lin, Z., Feng, L., Xiao-dong, S. 2017. *Recent progress and future prospect of digital soil mapping: A review*. Journal of Integrative Agriculture, 16(12): 2871–2885.
[16] Gee, G.W., Bauder J.W., 1986. Particle-size analysis. In: A. Klute (ed.), *Methods of soil analysis*. Part 1. *Physical and mineralogical methods*. 2nd ed., SSSA Book Series 5. ASA and SSSA, Madison, pp. 383–411, DOI: 10.1108/09593840110411167.
[17] Grunwald, S., 2006. *Environmental Soil-Landscape Modeling. Geographic Information Technologies and Pedometrics*, Taylor and Francis, Boca Raton–London–New York.
[18] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. *The WEKA data mining software: an update*. ACM SIGKDD Explore News, 11: 10–18.
[19] Hengl, T., Huvelink, G.B.M., Stein, A., 2004. *A generic framework for spatial prediction of soil variables based on regression-kriging*. Geoderma, 120: 75–93.
[20] Hengl, T., Toomanian, N., Reuter, H.I., Malakouti, M.J., 2007. *Methods to interpolate soil categorical variables from profile observations: Lessons from Iran*. Geoderma, 140: 417–427.
[21] Hsu, S.H., Hsieh, J.J.P., Chih, T.C., Hsu, K.C., 2009. *A two-stage architecture for stock price forecasting by integrating self-organizing map and support vector regression*. Journal of Expert Systems with Applications, 36: 7947–7951.
[22] Jafari, A., Finke, P., Wauw, J., Ayoubi, S., Khademi, H., 2012. *Spatial prediction of USDA-great soil groups in the arid Zarand region, Iran: Comparing logistic regression approaches to predict diagnostic horizons and soil types.* European Journal of Soil Science, 63: 284–298.
[23] Jafari, S., Baghdadi, G., Hashemi Golpayegani, S.M.R., Towhidkhah, F., Gharibzadeh, S., 2013. *Is attention deficit hyperactivity disorder a kind of intermittent chaos?* Neuropsychiatry and Clinical Neurosciences. 25, E02.
[24] Karaboga, D., Bahriye, A., 2009. *A comparative study of Artificial Bee Colony algorithm*. Applied Mathematics and Computing, 214: 108–132.
[25] Keller, T., Dexter, A.R., 2012. *Plastic limits of agricultural soils as functions of soil texture and organic matter content*. Soil Research, 50: 7–17.
[26] Keshavarz, A., Ghasemian Yazdi, V.H., 2004. *A fast algorithm based on support vector machine for classification of hyperspectral images using spatial correlation* (in Persian). Computer, Engineering and Electrical of Iran, 3: 37–44.
[27] Kim, J., Hwang, M., Jeong, D.H., Jung, H., 2012. *Technology trends analysis and forecasting application based on decision tree and statistical feature analysis*. Expert Systems with Applications, 39: 12618–12625.
[28] Kovačević, M., Bajat, B., Gajić, B., 2010. *Soil type classification and estimation of soil properties using support vector machines*. Geoderma. 154: 340–347.
[29] Kuhn, M., Johnson, K., 2013. *Applied Predictive Modeling*. Springer Science+Business, New York, XIII, 600.
[30] Lagacherie, P., McBratney, A.B., 2007. *Spatial soil information systems and spatial soil inference systems: perspectives for digital soil mapping. Digital soil mapping: An introductory perspective*. Developments in Soil Science, 31: 3–22.
[31] Levenberg, K., 1944. *A method for the solution of certain non-linear problems in least squares*. Quarterly of Applied Mathematics, 2: 164–168.
[32] Liao, K., Xu, S., Wu, J., Zhu, Q., 2013. *Spatial estimation of surface soil texture using remote sensing data*. Soil Science and Plant Nutrition, 59: 488–500.
[33] Marchetti, A., Piccini, C., Santucci, S., Chiuchiarelli, I., Francaviglia, R., 2011. *Simulation of soil types in Teramo province (Central Italy) with terrain parameters and remote sensing data*. CATENA, 85: 267–274.

[34] Marquardt, D.W., 1963. *An algorithm for least-squares estimation of nonlinear parameters*. Journal of the Society for Industrial and Applied Mathematics, 11(2): 431–441.
[35] Mena, J., 1999. *Data Mining Your Website*, Digital Press, Woburn.
[36] Metternicht, G.I., Zinck, J.A., 2003. *Remote sensing of soil salinity: Potentials and constraints*. Remote Sensing Environment, 85: 1–20.
[37] Minasny, B., McBratney, A.B., 2006. *A conditioned Latin hypercube method for sampling in the presence of ancillary information*. Computers and Geosciences, 32: 1378–1388.
[38] Mirkhani, R., Saadat, M., Shaban-Poorshahrestani, M., Aria, P., Yegane, M., 2006. *Evaluating of soil stability using readily available of soil properties* (in Persian). Water and Soil Science, 21(2): 201–207.
[39] Moonjun, R., Farshad, A., Shrestha, D.P., Vaiphasa, C., 2010. *Artificial neural network and decision tree in predictive soil mapping of Hoi Rin sub-watershed, Thailand*. Digital Soil Mapping: Bridging Research, Environmental Application and Operation, 2: 151–164.
[40] Moradi, S., 2013. *Effects of CEC on Atterberg limits and plastic index in different soil textures*. International Journal of Agronomy and Plant Production, 4: 2111–2118.
[41] Mukhlisin, M., Rahman, A., 2014. *Prediction of Atterberg limits via ANN and ANFIS: A comparison*. Recent Advances in Environmental Science and Geoscience, 69–74.
[42] Oreski, S., Oreski, D., Oreski, G., 2012. *Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment*. Expert Systems with Applications, 39(16): 12605–12617.
[43] Pahlavan Rad, M.R., Toomanian, N., Khormali, F., Brungard, C.W., Komaki, C.K., Bogaert, P., 2014. *Updating soil survey maps using random forest and conditioned Latin hypercube sampling in the loess derived soils of northern Iran*. Geoderma, 232(234): 97–106.
[44] Panigrahi, B.K., Yuhui, S., Meng-Hiot, L., 2011. *Handbook of Swarm Intelligence*, Springer-Verlag, Berlin–Heidelberg.
[45] Pao, Y.-H., 1989. *Adaptive Pattern Recognition and Neural Networks*, Addison-Wesley Longman Publishing, Reading, MA.
[46] Poggio, L., Gimona, A., Brewer, M., 2013. *Regional scale mapping of soil properties and their uncertainty with a large number of satellite derived covariates*. Geoderma, 209(210): 1–14.
[47] Quinlan, J.R., 2001. *Cubist: An Informal Tutorial*, http://www.rulequest.com
[48] Rahimnia, R., Heidari Bani, V.D., 2010. *Effect of soil plasticity index on tensile and compressive strength of stabilized brick with cement for using protection of adobe structures* (in Persian). Repair, History Cultural of Monuments and Tissues, 2(1): 91–102.
[49] Saikia, A., Baruah, D., Das, K., Rabha, H.J., Dutta, A., Saharia, A., 2017. *Predicting compaction characteristics of fine-grained soils in terms of Atterberg limits*. International Journal of Geosynthetics and Ground Engineering, 3(18): 1–9.
[50] Scull, P., Franklin, J., Chadwick, O.A., McArthur, D., 2003. *Predictive soil mapping: A review*. Progress in Physical Geography, 27: 171–197.
[51] Sherzoy, M.M., 2017. *Atterberg limits prediction comparing SVM with ANFIS model*. Geoscience, Engineering, Environment and Technology, 2(1): 20–30.
[52] Stum, A.K., Boettinger, J.L., White, M.A., Ramsey, R.D., 2010. *Random forests applied as a soil spatial predictive model in arid Utah*. Digital Soil Mapping. Part of the progress in soil science book series, 2: 179–190.
[53] Taghizadeh-Mehrjardi, R., Nabiollahi, K., Kerry, R., 2016. *Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran*. Geoderma, 266: 98–110.
[54] Taghizadeh-Mehrjardi, R., Neupane, R., Sood, K., Kumar, S., 2017. *Artificial bee colony feature selection algorithm combined with machine learning algorithms to predict vertical and lateral distribution of soil organic matter in South Dakota, USA*. Carbon Management, 8: 277–291.

[55] Tol, J.J.V., Dzene, A.R., Roux, P.A.L.L., Schall, R., 2016. *Pedotransfer functions to predict Atterberg limits for South African soils using measured and morphological properties*. Soil Use and Management, 32: 635–643.

[56] Yildirim, B., Gunaydin, O., 2011. *Estimation of California bearing ratio by using soft computing systems*. Expert Systems with Applications, 38: 6381–6391.

[57] Zentar, R., Abriak, N.E., Dubois, V., 2009. *Effects of salts and organic matter on Atterberg limits of dredged marine sediments*. Applied Clay Science, 42: 391–397.

[58] Zeraatpisheh, M., Ayoubi, S., Jafari, A., Finke, P., 2017. *Comparing the efficiency of digital and conventional soil mapping to predict soil types in a semi-arid region in Iran*. Geomorphology, 285: 186–204.

[59] Zolfaghari, Z., Mosaddeghi, M.R., Ayoubi, S., 2015. *ANN-based pedotransfer and soil spatial prediction functions for predicting Atterberg consistency limits and indices from easily available properties at the watershed scale in western Iran*. Soil Use and Management, 31(1): 142–154.