Journal of Hydrology 549 (2017) 592–602

Contents lists available at ScienceDirect

Journal of Hydrology

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

Research papers

Modeling soil bulk density through a complete data scanning procedure: Heuristic alternatives

Jalal Shiri [a], Ali Keshavarzi [b,*], Ozgur Kisi [c], Sepideh Karimi [a], Ursula Iturraran-Viveros [d]

[a] Water Engineering Department, Faculty of Agriculture, University of Tabriz, Tabriz, Iran
[b] Laboratory of Remote Sensing and GIS, Department of Soil Science, University of Tehran, P.O. Box: 4111, Karaj 31587-77871, Iran
[c] Center for Interdisciplinary Research, International Black Sea University, Tbilisi, Georgia
[d] Departamento de Matematicas, Facultad de Ciencias, Universidad Nacional Autonoma de Mexico, Circuito Escolar, Cd. Universitaria, Coyoacán 04510, Ciudad de México, Mexico

ARTICLE INFO

Article history:
Received 28 February 2017
Received in revised form 13 April 2017
Accepted 15 April 2017
Available online 19 April 2017

This manuscript was handled by G. Syme, Editor-in-Chief

Keywords:
Heuristic models
K-fold testing
Pedotransfer functions
Soil bulk density

ABSTRACT

Soil bulk density (BD) is very important factor in land drainage and reclamation, irrigation scheduling (for estimating the soil volumetric water content), and assessing soil carbon and nutrient stock as well as determining the pollutant mass balance in soils. Numerous pedotransfer functions have been suggested so far to relate the soil BD values to soil parameters (e.g. soil separates, carbon content, etc). The present paper aims at simulating soil BD using easily measured soil variables through heuristic gene expression programming (GEP), neural networks (NN), random forest (RF), support vector machine (SVM), and boosted regression trees (BT) techniques. The statistical Gamma test was utilized to identify the most influential soil parameters on BD. The applied models were assessed through k-fold testing where all the available data patterns were involved in the both training and testing stages, which provide an accurate assessment of the models accuracy. Some existing pedotransfer functions were also applied and compared with the heuristic models. The obtained results revealed that the heuristic GEP model outperformed the other applied models globally and per test stage. Nevertheless, the performance accuracy of the applied heuristic models was much better than those of the applied pedotransfer functions. Using k-fold testing provides a more-in-detail judgment of the models.

1. Introduction

Soil bulk density (BD) is defined as the dry weight of soil per soil volume (which includes the soil particles as well as the soil pores). It is an important factor in land drainage and reclamation because it is an indicator of drainage characteristics (Arya and Paris, 1981; Braun and Kruijne, 1994), and determines whether there are impermeable barriers in the soil, which can deteriorate the drainage and root penetration conditions (Lampurlanes and Cantero-Martinez, 2003). In irrigation scheduling, BD can be utilized to estimate the soil volumetric water content which is an important parameter for controlling optimum irrigation (Howell and Meron, 2007). BD is also an essential factor for assessing soil carbon and nutrient stock (Ellert and Bettany, 1995), determining pollutant mass balance in soil, and determining the soils' packing structure in soil classification issues (Dexter, 1988). It also affects the soil biomass productivity and environment quality (Lal and Kimble, 2001).

Recently, the soil processes simulating models have been developed for improving the existing knowledge on important soil processes as well as evaluating the agricultural and environmental problems (Minasny and McBratney, 2002). On the other hand, soil properties continuously vary across the landscape. Nevertheless, there are usually limited numbers of direct observations (visual inspection, sampling and observation) in the field, which make it difficult to determine some soil properties directly (Heuvelink and Webster, 2001). So, numerous investigations have been conducted to relate some soil properties to easily measured available soil characteristics (e.g. particle-size distribution, organic matter or organic C content, BD, porosity, etc). Such relationships are called as pedotransfer functions (Mermoud and Xu, 2006).

The common methods of developing pedotransfer functions are multi variate-linear regression and artificial intelligence-based models (Schaap and Leij, 1998). Among others, Jalabert et al. (2010) applied boosted regression trees (BT) for estimating forest soil BD and found that the variations in forest soil BD magnitudes are largely affected by organic carbon (OC) content, followed by tree species, the coarse fragment content, parent material and depth of sampling. Ghehi et al. (2012) applied k-nearest neighbor

* Corresponding author.
E-mail address: alikeshavarzi@ut.ac.ir (A. Keshavarzi).

http://dx.doi.org/10.1016/j.jhydrol.2017.04.035
0022-1694/

and boosted trees methods for predicting top soil BD of a tropical mountain forest soils and found the OC as the most influential parameter on BD. Al-Qinna and Jaber (2013) used different techniques inducing linear/nonlinear regression and neural networks (NN) for estimating BD using data from an arid environment in Jordan and found NN as the best model among other applied models. Botula et al. (2015) compared multivariate linear regression and k-nearest neighbor methods for predicting BD for the soils of Central Africa. They used independent soil samples for further testing of the developed models and observed substantial differences between the observed and predicted BD magnitudes, which imply the difficulties in generalizing the pedotransfer functions for its estimation. Rodríguez-Lado et al. (2015) compared random forest (RF), linear regression and NN in estimating BD and found the RF as the most accurate model among the applied models. They also confirmed that the soil BD in the studied area was mainly influenced by the soil OC. Xiangsheng et al. (2016) compared linear regression and NN models to develop pedotransfer functions for BD estimation and confirmed the superiority of NN, which confirms the conclusions obtained by Patil and Chaturvedi (2012) regarding the NN superiority.

The presented studies have used different methods for determining the input variables of the applied models. In heuristic models, feature selection (or variable choice) is the way toward choosing a subset of important elements for utilizing in model building. The focal preface when utilizing a feature selection strategy is that the data involves various features that are unessential, and can in this manner be expelled without bringing about much loss of information. Unessential features are two particular concepts, since one pertinent feature might be repetitive in the asset of another appropriate feature with which it is highly correlated. There are two types of variable selection methods: wrapper and filter types (Pfleger et al., 1994). Wrapper techniques assess variable subsets which permit to identify the probable interactions between variables (see Phuong et al., 2005). Filter methods (e.g. Gamma test, which was used in the current study) evaluate the set of variables directly from the data itself. Model construction based on wrapper methods is rather inflexible though it may be advantageous in some situations where model selection is integrated together with the variable selection method.

Nonetheless, most of the existing literatures have used a single data set assignment for developing and testing the applied regression and heuristic models, where a part of available patterns are used for training the models, then the obtained models are tested using the rest of available patterns. Meanwhile, some studies have tried to test the obtained models using data outside the studied region as discussed by Botula et al. (2015). However, the present paper focused on developing pedotransfer functions of soil BD estimation by using gene expression programming (GEP), neural networks (NN), random forest (RF), support vector machine (SVM) and boosted regression trees (BT) techniques assessed through a k-fold testing. So, a complete data set scanning was conducted so that all available patterns can participate in train-test phases, which will avoid obtaining partially valid conclusions (that might be achieved using single data set assignment). A comparison was also made between the results of these models and those obtained through using the previously suggested regression-based pedotransfer functions.

2. Study area and used data

Data used in the present study were gathered in Mohr plain, Fars province, located in Southwestern Iran (Fig. 1), between the latitudes of 27° 25′ N to 27° 59′ N and longitudes of 52° 21′ E to 53° 05′ E, with an area about 1900 km^2. The area altitude varies from 282 m to 1780 m, while the main topographic elevation ranges over 1031 m (above sea level). The dominant soil types include Lithic Leptosols, Gypsic Regosols, Haplic Calsisols, Calcaric Cambisols and Calcaric Solonchaks, which cover mountains, hilly land, plains and colluvial fans.

The main land uses practiced in the study area are pastures and irrigated farming across the Mehran River. A simple random sampling scheme was designed using ArcGIS 10.2.2 software for an appropriate determination of soil sampling areas to consider spatial varieties of the parameters affecting the BD in the study region. A total of 250 soil samples were obtained from two-first vertical depths (0–30 and 30–60 cm depth) of 125 representative soil profiles. Depths were assigned to a soil textural class determined by the substances of clay, silt, and sand, as indicated by the USDA textural triangle (Schoeneberger et al., 2002). The soil texture classes are illustrated in Fig. 2. Table 1 summarizes the statistical characteristics of the applied data. From Table 1 it is seen that soil BD has a wide variability range with the minimum and maximum values of 1.134 and 1.964 g/cm^3, respectively. The variations of the other applied parameters (except pH) are also strongly high representing high variability class according to the coefficient of variation values (Adrover et al., 2012).

The sampling sites were designed to cover equally the entire area and to incorporate different soil and land use types. The collected disturbed soil samples were air dried, crushed and sieved using 2 mm sieve size. Large plant material and pebbles were separated and discarded. Soil organic carbon (OC) content was obtained by the Walkley–Black technique with dichromate extraction and titrimetric quantization (Nelson and Sommers, 1986). Rates of clay (<0.002 mm), silt (0.002–0.05 mm), and sand (0.05–2 mm) particles were measured via hydrometer method (Gee and Bauder, 1986). Soil pH was measured in saturated paste extract utilizing a digital pH-meter (Thomas, 1996) and calcium carbonate equivalent (CCE) was obtained by the back-titration technique (Nelson, 1982). Finally, the clod method (Blake and Hartge, 1986) was utilized for determining BD with triple replications.

2.1. Input selection

The Gamma test is a filter method that assesses the significance of the variables directly from the data. A rigorous mathematical proof of the method is presented in the work by Evans (2002). The Gamma test was originally conceived as a method of estimating the variance of the residual of a model. It has also been used successfully in the estimation of optimum embedding dimension and lag of chaotic systems (Tsui et al., 2002) and the assessment of the quality of data (Jones et al., 2002). The Gamma test can be used to predict the reliability of models before the heuristic models training phase begins saving lots of time. In this paper, we seek to estimate the BD from 6 input variables: Clay, Silt, Sand, CCE, OC and pH. Therefore, the basic relationship of the system under investigation is of the accompanying form:

$$Bulk\quad density = f(\mathrm{Clay},\ \mathrm{Silt},\ \mathrm{Sand},\ \mathrm{CCE},\ \mathrm{OC},\ \mathrm{pH}) + r, \tag{1}$$

where f is a smooth function and r is a random variable that indicates noise. The domain of conceivable models is limited to the class of smooth capacities which have limited first partial derivatives. The Gamma statistic Γ is the estimate of that part of the output variance that can't be represented by a smooth data model.

Taking into account that we consider 6 possible input variables for the model building then the number of possible combinations is $2^6 - 1 = 63$. Among these possible input combinations, we might want to choose those which have the least Γ statistic. The Gamma statistic Γ also gives a lower bound for the mean square error (MSE). When training heuristic models, this implies that we should stop the training process when the MSE reaches the Gamma

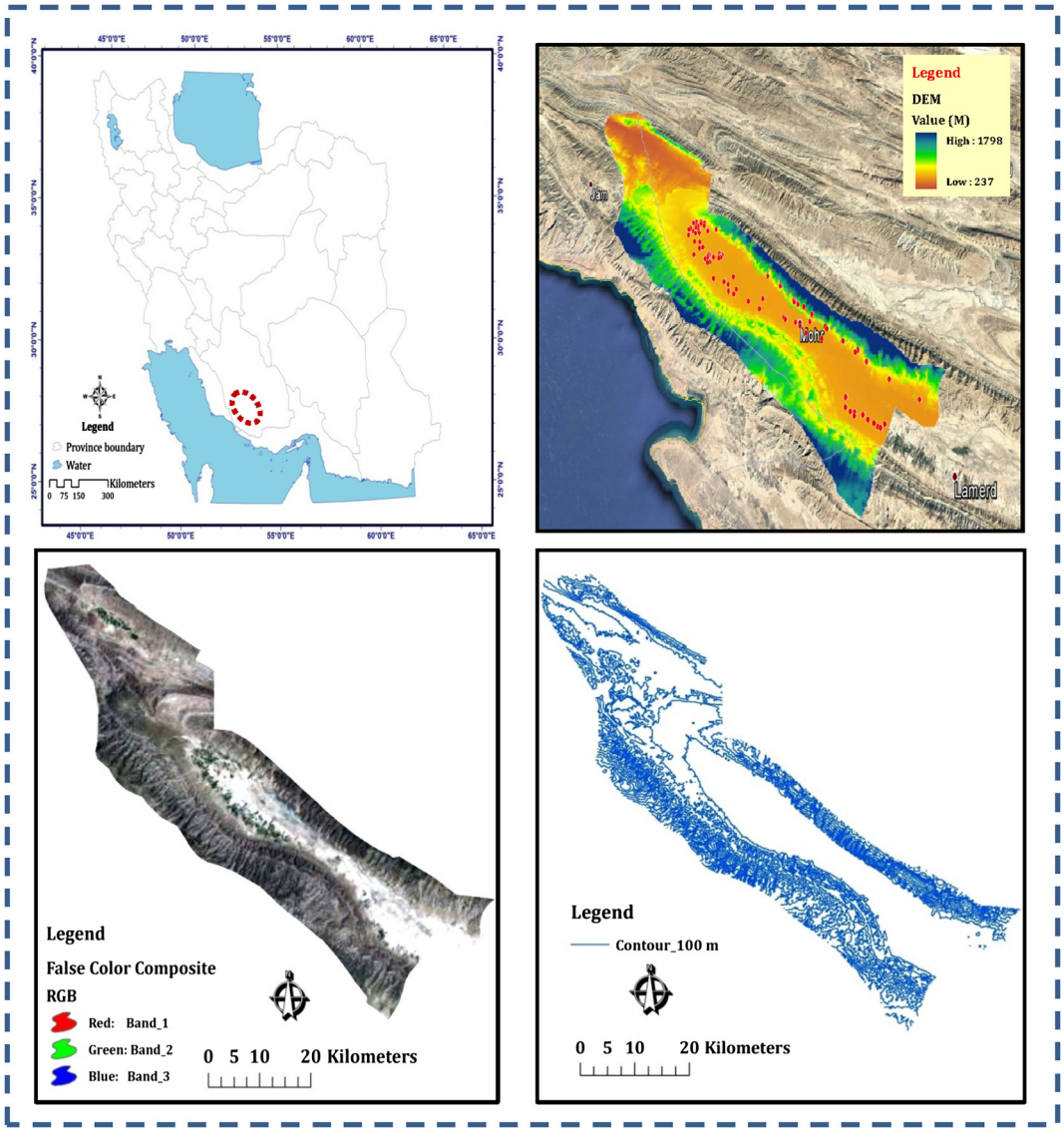

Fig. 1. Location of study area showing sampling points, false color composite of google earth imagery and contour lines with 100 m interval (Mohr region, Iran).

statistic Γ, avoiding overfitting. Accordingly, if we train these models with a combination that has a large Gamma statistic Γ, the performance of this model will deteriorate since we would not be able to lower the MSE below the Gamma statistic Γ. Summarizing, it was found that the input combination 101111 indicating Clay, Sand, CCE, OC and pH variables has the best influence to soil BD with the lowest Gamma statistic while the input combination of Clay, Silt, CCE and pH has the lowest influence.

2.2. Data splitting

A cross-validation technique (k-fold testing) is used for splitting the input-output matrices in the present study. So a complete data scanning, in which a portion of data is held out each time, was carried out by dividing the whole patterns into 10 sub-portions. Therefore, all the available patterns of the input-output matrix were used in both the train and test stages and no any pattern was remained unseen. This would reduce the risk of model overfitting as well as getting partially valid conclusions which might be drawn down using traditional data management scenarios (Marti et al., 2013; Shiri et al., 2014a). Given that the GEP, NN, SVM, RF and BT models were applied in the present study, a total of 50 models (5 techniques∗10 configurations) were established. A similar procedure was also repeated for the applied regression-based pedotransfer functions, so a total of 40 models (4 functions∗10 configurations) were constructed.

Two statistical criteria were applied here for assessing the models performance, namely, the mean absolute error (*MAE*), and the scatter index (*SI*), expressions for which are:

$$MAE = \frac{\sum_{i=1}^{n} |BD_{io} - BD_{im}|}{n} \tag{2}$$

$$SI = \frac{RMSE}{\overline{BD_o}} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n} (BD_{io} - BD_{im})^2}}{\overline{BD_o}} \tag{3}$$

In the above equations, BD_o is the observed *BD* value at the *i*-th observation step, BD_m is the corresponding simulated *BD* value, *n* represents the number of data patterns, $\overline{BD_o}$ denotes the mean observed value and *RMSE* is the root mean square errors. These

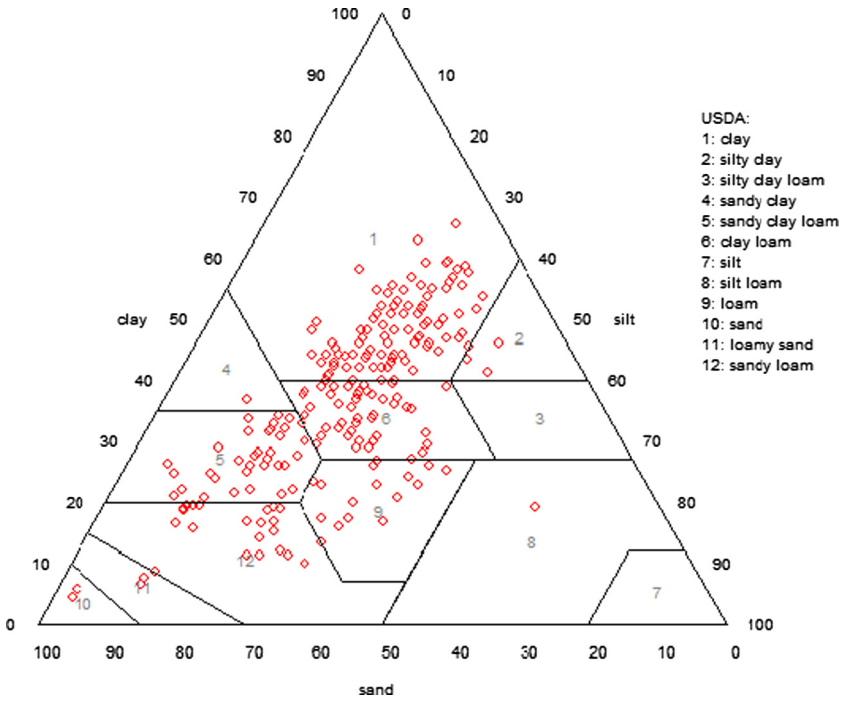

Fig. 2. Soil texture classes in study area (red points showing the samples).

Table 1
Statistical characteristics of the used data.

	Clay (%)	Silt (%)	Sand (%)	CCE (%)	OC (%)	pH (-)	BD (g/cm^3)
Maximum	65.440	62.800	92.720	48.700	1.330	8.320	1.964
Minimum	4.280	3.000	6.560	3.200	0.094	7.420	1.134
Mean	36.048	26.534	37.410	16.513	0.587	7.993	1.424
Standard deviation	13.114	8.284	17.001	6.691	0.274	0.168	0.131
Coefficient of variation	0.364	0.312	0.454	0.405	0.467	0.021	0.092
Skewness	−0.213	0.168	0.551	0.925	0.370	−0.504	0.262
Kurtosis	−0.724	1.227	−0.043	2.223	−0.411	−0.116	0.352

Note: CCE: Calcium carbonate equivalent; OC: Soil organic carbon; pH: Soil acidity; BD: Soil bulk density.

indicators were computed for the complete series for data set (pooling the simulations of each data set) as well as split up for each test stage (one-tenth of the data matrix).

3. Methods

3.1. Gene expression programming (GEP)

Genetic Programming (GP) (Koza, 1992), utilizes a "parse tree" structure for searching its solutions. GP has the capability for deriving a set of explicit formulations that govern the studied phenomenon, to describe the relationship(s) between the input-target variables using different operators. Gene expression programming (GEP) is similar to GP, in a way that selects the best governing formulations based on fitness values and introduces genetic variation using a unique or various genetic operators (Ferreira, 2006). One of the advantages of GP (i.e. GEP) over other heuristic techniques (e.g. NF, NN and SVM) is in giving explicit expression of the input-target relationship. Further details about modeling process with GEP can be read in e.g. Ferreira (2006).

In the present work, GeneXpro program (Ferreira, 2006) was used for GEP simulations. The procedure for GEP-based modeling of soil BD (target variable) as a function of soil parameters (input variables) is as follows (Ferreira, 2001):

i) Selection of the fitness function: different absolute- and relative-error based fitness functions were utilized and evaluated in modeling soil BD (Table 2). The *SI* values presented in Table 2 clearly show that the root relative squared error (*RRSE*) is the best fitness function among others. In some related literature (Kisi et al., 2012, 2013; Kisi and Shiri, 2012), the *RRSE* was also found to be the best fitness function for the optimal GEP model among others.

Table 2
Preliminary investigation on GEP fitness function using *SI*.

Fitness function based on the absolute error	*SI*	Fitness function based on the relative error	*SI*
Absolute error with selection range*	0.081	Relative error with selection range*	0.081
Absolute/hits	0.764	Relative/hits	0.772
Mean squared error (MSE)	0.080	r-MSE*	0.083
Root mean squared error (RMSE)	0.080	r-RMSE	0.080
Mean absolute error (MAE)	0.084	r-MAE	0.0845
Relative squared error (RSE)	0.080	r-RSE	0.081
Root relative squared error (RRSE)	0.077	r-RRSE	0.080
Relative absolute error (RAE)	0.084	r-RAE	0.086

* Selection range was selected as maximum fitness; r: denotes relative error-based functions.

ii) Selection of the input parameters and function sets. The input parameters (terminals) comprise clay, sand, OC, CEC and pH. Various GEP function sets as listed in Table 3 were evaluated to select the proper set. Analyzing the *SI* values presented in Table 3 shows that the function set F5 with addition linking function outperforms the other evaluated sets. However, for avoiding nested function production, the function set F4 was selected to provide less complex expressions. It is worth noting that the primitives functions which can be used by genetic programming family are +,−,*,/ according to recommendations given by Koza (1992). However, the applied sets are the combinations of the functions which have been found to perform better than the other in different literature (e.g. Shiri et al., 2012, 2017).

iii) Selecting chromosomal architecture: different values of head length and genes per chromosomes were evaluated and compared in Fig. 3. It is clear that the head size = 8 and number of genes = 3 produce the optimum results. It is apparent from the figure that the *SI* decreases by increasing head size up to 8 and then it increases. In case of gene numbers, the variations in model's accuracy beyond 3 genes are negligible, so this point was selected for avoiding more complexity in GEP trees.

iv) Selection of the genetic operators. GeneXpro default operators were selected in this step as have been advised by literature (Number of chromosomes: 30, head size: 8, number of genes: 3, linking function: addition, fitness function error type: root relative squared error, mutation rate: 0.044, inversion rate: 0.1, one point recombination rate: 0.3, two point recombination rate: 0.3, gene recombination rate: 0.1, gene transposition rate: 0.1, insertion sequence transposition rate: 0.1, root insertion sequence transposition: 0.1) (Shiri and Kisi, 2012).

3.2. Neural networks (NN)

Neural networks are parallel information-processing systems that have been originally designed for simulating the performance of a biological neural system. The commonly used structure of NN is consisted of the input, hidden, and output layers that is known as multilayer perceptron (MLP) (Fausset, 1994). A three-layer feed-forward network was used in the current paper with different transfer functions in the hidden and output layers. The hidden-layer-node numbers of the models were determined via a trial and error process. At each training process, 100 networks were examined and the optimum structure of each case (transfer functions) was chosen. The minimum and maximum values of weight decay in hidden layer were found as 0.0001 and 0.002. Three different NN models with different activation functions and it was found that the NN model with Tanh and Exponential activation functions in hidden and output neurons comprising 31 neurons in hidden layer gave the most accurate results.

Table 3
Preliminary investigation on GEP function set.

	Definition	*SI*
F1	$\{+, -, \times, \div\}$	0.081
F2	$\{+, -, \times, \div, \sqrt{}, x^2\}$	0.080
F3	$\{+, -, \times, \div, \sqrt{}, Power, Lnx, Logx, e^x, 10^x\}$	0.080
F4	$\{+, -, \times, \div, \sqrt{}, Arctgx\}$	**0.077**
F5	$\{+, -, \times, \div, \sqrt[3]{}, \sqrt{}, \ln, e^x, x^2, x^3, \sin x, \cos x, Arctgx\}$	0.077
1	Addition	**0.077**
2	Multiplication	0.080
3	Subtraction	0.080
4	Division	0.083

3.3. Random Forest technique (RF)

Random Forests (RF) is a group learning algorithm that manages high-dimension regression problems. It is a tree-based group method, where all trees are dependent of a collection of random variables, and the forest is grown from many regression trees put together and from a group (Breiman, 2001). The final decision is resulted via averaging the output, after fitting single trees in ensemble (bagging procedure). The bias of the bagged trees is the same as that of the single trees, while the variance is decreased by reduction in the correlation between trees (Hastie et al., 2009).

Different numbers of trees were evaluated to select the optimum random forest method as has been depicted in Fig. 4. The figure clearly shows that the variations in error magnitude is more obvious for tree numbers lower than 100, while the error magnitudes monotonously fluctuate for trees more than 100. Finally, the optimum tree number was selected as 158, as it produces the lowest average squared error among others. In the stopping conditions, 10 cycles were found to be the best for calculating the mean error, iteratively. Also the percentage decrease in training error was found to be as 5% by trial and error. Minimum child node size to stop (which controls the smallest permissible number in a child node, for a split to be applied) and the maximum number of levels (the depth of the tree as measured from the root node) were chosen as 5, and 10, respectively (based on trial and error).

3.4. Boosted regression trees (BT)

BT has an algorithm that incorporates tree based methods with boosting, a method with origins in the machine learning field that may be evaluated as an improved form of regression (Friedman et al., 2000). It is an ensemble strategy that incorporates the ability of boosting and regression trees algorithms for fitting statistical models that are basically dissimilar to classical methods that try to fit a single parsimonious model. The BT model can be comprehended as an additive regression, in which individual terms are basic trees, fitted in a forward, stage wise style. Boosted regression tree combines major benefits of tree-based methods, taking care of various types of predictor parameters and accommodating missing information (data). It has no requirement for prior data transformation or removing outliers, can fit complicated non-linear relationship, and naturally processes interaction impacts between predictors. Fitting numerous trees in BT defeats the greatest disadvantage (poor prediction accuracy) of single tree models. In spite of the fact that BRT is complex, it can be outlined in ways that provide powerful ecological understanding, and its prediction accuracy is better than the most conventional Modeling approaches (Elith et al., 2008; Toming et al., 2016). As a recently created class of analytical calculation, BT has not yet observed broad application the hydrological sciences, in spite of the fact that uses of this data intensive strategy have expanded in the last decade (e.g., Tisseuil et al., 2010; Erdal and Karakurt, 2013; Rice et al., 2015, 2016). Analytically, BT regularization includes together optimizing the quantity of trees, which are naturally tuned with inner cross-validation, learning rate (decides the contribution of every tree to the developed final model), bag fraction and node numbers in a single tree (tree complexity) which are controlled via trial and error (França and Cabral, 2015).

Boosted regression tree can be obtained utilizing a few control parameters for example, the node number in a tree, learning rate, the percentage of data chosen at every step, and the average tree numbers included in the ensemble forest (ntrees). As reported by Elith et al. (2008), a decrease in learning rate increases the ntrees value required, and a smaller learning rate value (and a larger ntrees value) is generally desirable, conditional on the observation number and the available time for calculation (Rodrigues and Riva,

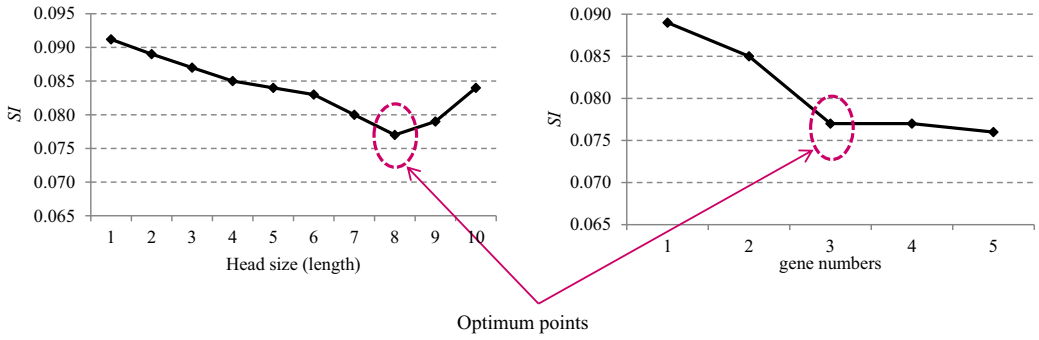

Fig. 3. *SI* values corresponded to different head length and gene numbers.

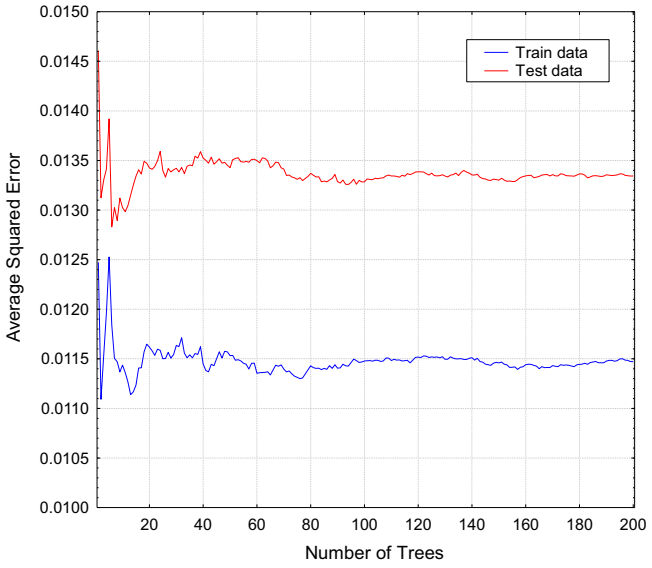

Fig. 4. Evaluating different tree numbers in RF method.

2014). The specific weight with which the consecutive simple trees are added into the estimation model (referred to as the learning rate) is usually a constant, for which the values of 0.1 or less usually provide better models (Friedman et al., 2000; Hastie et al., 2016). Number of additive trees (the number of simple regression trees to be computed in successive boosting steps) was selected as 200 based on trial and error. Different numbers of seeds for random number generator was examined and the optimum results were obtained by fixing this constant as 1. Also the minimum child node size (which is utilized for controlling the smallest permissible number in a child node) was found to be as 1, iteratively. Finally, maximum number of levels (depth of the tree as measured from the root node) was found to be as 10 for the perfect fit.

3.5. Support vector machine (SVM)

Support vector machine (SVM) is one of the novel soft computing algorithms which have been environmental issues. It majorly finds lots of utilizations in forecasting, pattern recognition, regression analysis and classification. Firstly developed by Vapnik et al. (1997), SVM is based on statistical machine learning procedure and structural risk minimization. It minimizes the upper bound generalization error instead of local training error, providing SVM a greater ability to generalize, which is the goal in statistical learning (Vapnik et al., 1997; Gunn, 1998). Further details on the application of SVM can be found e.g. in Vapnik et al. (1997).

Here the regression-SVM type1 is applied as it has demonstrated higher accuracy in previous studies (Shiri et al., 2014b). The difference between two types of regression-SVM belongs to their error functions (Vapnik et al., 1997). Consequently, using trial and error process, SVM constants were chosen as 10 (capacity) and 0.15 (epsilon). Different linear, sigmoid, polynomial and radial basis functions were examined as SVM kernel functions, from which the radial basis function kernel (Gamma = 0.25) provided much better results than the other kernels. Finally, maximum number of iterations was found to be 1000 (iteratively) and the models were stopped at error values of 0.005.

3.6. Existing pedotransfer functions

Among different existing pedotransfer functions, as used by De Vos et al. (2005) and Jalabert et al. (2010) as follows:

- Jeffrey (1970) analyzed the soil BD variations in relation to the soil organic matter (OM) using eighty samples of uncultivated soils and stated that OM is the main parameter affecting the BD in uncultivated soils (e.g. forest soils). So, a regression-based expression of BD relation with OM was suggested as:

$$BD = a + b\log_{10}(OM) \tag{4}$$

- Manrique and Jones (1991) developed a regression equation for estimating soil *BD* (at −33 kPa moisture content) as follows:

$$BD = a + b\sqrt{OC} \tag{5}$$

Adams (1973) argued that organic matter would create an expansion in soil mineral matter volume which equals its own bulk volume. Based on this research, the effect of soil structure in mineral agricultural soils would affect the involvement of soil organic matter on soil BD. Using different soil samples of eluvial soil layers, Adams (1973) proposed the following regression-based equations for *BD* estimation:

$$BD = \frac{100}{\{(OM/a) + [(100 - OM)/b]\}} \tag{6}$$

Federer (1983) evaluated the relations between the BD and OM based on an earlier work of Curtis and Post (1964), using 130 core samples of four study sites in New England and developed the following equation for estimating soil *BD*:

$$Ln(BD) = a_0 + a_1.Ln(OM) + a_2[LN(OM)]^2 \tag{7}$$

where, BD denotes soil bulk density (g/cm3), OC is the soil organic carbon content (%), and OM stands for the estimation of soil organic matter (%) which can be considered as OM = 1.724∗OC (Jalabert et al., 2010). The regression constants (*a*, *b*, etc) might be computed through making a regression using the total available data. In the

present study, since the k-fold procedure was applied, the coefficients were determined for all train-test sets.

4. Results and discussions

Global error statistics of the GEP, RF, BT and SVM models are listed in Table 4. These values are corresponding to the global performance of the models which are computed using all observed-simulated patterns. From Table 4, it is clearly seen that GEP model outperforms the other applied models with the lowest *SI* (0.069) and *MAE* (0.076) values. BT and SVM are ranked as the second models with respect to *SI* while the RF has a lower *MAE* (0.078) than the BT and SVM. However, the overall discrepancy between the models performance accuracy is small (0.004 and 0.006, *SI* and *MAE* difference between their maximum and minimum values, respectively). The table also presents the *t*-test results for the applied models which verify the robustness (the degree of differences between the observed values and the estimates of the applied model). Test was set at a significant level of 95%. Analyzing the statistics confirms that the GEP models surpasses the other applied models with the lowest *t*-statistics and the highest results at significance level. This means that the GEP model has a higher similarity between the observed soil BD and estimates than the other methods.

Comparing the performance of the heuristic models with those obtained by using the existing pedotransfer functions (Table 4) shows that, except the pedotransfer function suggested by Adams (1973), the applied functions cannot simulate the soil BD values with higher accuracy. As mentioned, these functions were assessed through k-fold testing so all the available patterns were involved in training (calibrating) and testing stages for discovering the optimum regression coefficients (constants) values. Although both Jeffrey (1970) and Federer (1983) use the soil OM as an input variable to simulate BD, the way this variable incorporate to model the target phenomenon seems to be more effective, so the outcomes of these functions are no longer promising in this regard. So, the mathematical expression chosen for relating the soil parameters to the BD has a key role in mapping the nonlinear relations between the soil parameters and its BD.

Among the applied heuristic models, one of the advantages of genetic programming-based models is providing the mathematical expressions of the functional relationships between the input and target parameters. This is a black and white approach which depicts the interrelations among the influential parameters as well as the target variable. The obtained GEP-based models for simulating soil BD in the present study reads:

$$BD = -0.247OC.arctg\left[\frac{clay}{CCE+7.02216}\right] + \frac{OC.arctg(pH)}{CCE+10.505} + 1.53433 \quad (8)$$

The equation clearly shows that GEP has not picked up the soil sand as input parameter for modeling BD, while Gamma test has identified it as an influential input parameter. This might be explained through comparing the differences between the GEP and Gamma-test methods. Gamma test is a nonlinear approach which allows examining the nature of a hypothetical input-target relation within a numerical data-set. Nonetheless, the basic assumption with using Gamma test is that the input matrix includes some parameters which influence the studied target (here, BD). The second assumption is that the governing interrelation of the studied system comprises a smooth function as well as a random variable, for which the domains of possible models would be restricted by the class of smooth functions' classes that have bounded the first partial derivatives. On the other hand, GEP simultaneously develops the structure and constants of the formulas, the extent that the site-specific constants are embedded in the formulation -and whether these constants are similar or different than those at another station- would dictate the transferability of the function between stations (Deschaine, 2014). Nonetheless, parsimony pressure tool was used here to reduce the tree (program) size so that it does not over-fit the data by becoming excessively customized. There are different schools of thought on using parsi-

Table 4
Global error statistics of the applied models.

	GEP	NN	RF	BT	SVM	Jeffrey	Manrique-Jones	Adams	Federer
MAE (g/cm^3)	0.076	0.080	0.078	0.081	0.082	0.394	0.164	0.101	0.321
SI	0.069	0.074	0.073	0.071	0.071	0.293	0.141	0.100	0.245
	t-Statistic			Resultant significance level					
GEP	−0.160			0.872					
NN	0.546			0.579					
RF	0.557			0.577					
BT	0.547			0.565					
SVM	0.814			0.415					
Jeffery	4.521			0.0001					
Manrique-Jones	3.880			0.0003					
Adams	0.914			0.411					
Federer	8.547			< 0.0001					

Table 5
Correlations between soil parameters and errors of the applied models.

	Clay	Sand	CCE	OC	pH	BD
Clay	1.000	−0.880	0.120	0.348	−0.212	−0.318
Sand	−0.880	1.000	−0.012	−0.324	0.139	0.212
CCE	0.120	−0.012	1.000	−0.097	−0.097	0.034
OC	0.348	−0.324	−0.097	1.000	−0.417	−0.510
pH	−0.212	0.139	−0.097	−0.417	1.000	0.170
BD	−0.318	0.212	0.034	−0.510	0.170	1.000
GEP	−0.605	0.548	0.174	−0.930	0.404	0.537
NN	−0.346	0.307	−0.024	−0.622	0.179	0.611
SVM	−0.481	0.252	0.049	−0.757	0.195	0.608
RF	−0.444	0.394	0.058	−0.784	0.242	0.620
BT	−0.543	0.425	0.153	−0.645	0.248	0.606

mony, since it is a binding or forcing function on an otherwise freely (i.e. unconstrained with maximal degrees of freedom) search strategy (when the program is allowed to grow unrestricted) (Francone and Deschaine, 2004). GEP gives the highest weight (variable importance) to the OC and CCE. The RF and BT models however give the maximum weight to OC, while clay is ranked as the second important input. The results confirm the conclusions obtained by previous studies where they introduced the soil OC as the most influential parameter on soil BD (e.g. Jalabert et al., 2010; Ghehi et al., 2012). Table 5 sums up the correlations between the soil parameters and the errors of the applied models. From the table, it is seen that the highest correlation (among the other applied parameters) belongs to BD with soil OC (r = −0.510, moderate correlation) followed by the soil separates: clay (r = 0.318) and sand (r = 0.212). Soil pH and CCE present lower magnitudes of correlations (r = 0.170, and r = 0.034, respectively). This implies that the GEP results confirm the correlation analysis (CA) results and CA can also provide useful information for the influence of

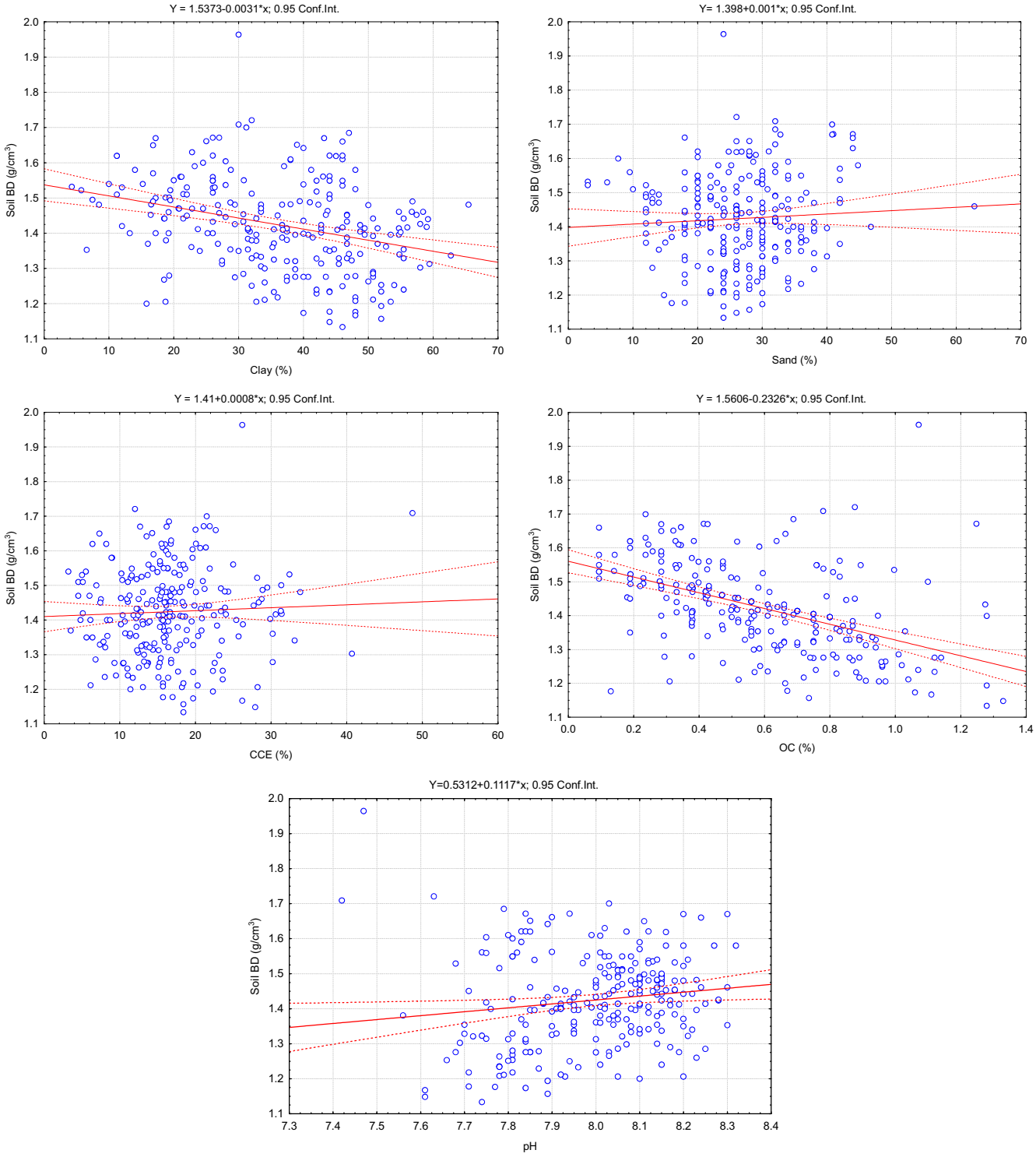

Fig. 5. Effect of different soil parameters on BD.

Table 6
Error statistics of the single-input models.

	GEP		NN		SVM		RF		BT		**Average**	
	SI	*MAE*	*SI*	*MAE*	*SI*	*MAE*	*SI*	*MAE*	*SI*	*MAE*	***SI***	***MAE***
Clay	0.090	0.100	0.099	0.110	0.097	0.100	0.098	0.104	0.102	0.105	**0.097**	**0.104**
Sand	0.090	0.098	0.098	0.115	0.091	0.100	0.095	0.103	0.097	0.104	**0.094**	**0.104**
CCE	0.091	0.099	0.091	0.099	0.091	0.099	0.094	0.103	0.095	0.103	**0.092**	**0.101**
OC	0.088	0.095	0.095	0.104	0.094	0.104	0.094	0.104	0.095	0.105	**0.093**	**0.102**
pH	0.112	0.110	0.116	0.115	0.112	0.114	0.128	0.116	0.134	0.121	**0.120**	**0.115**
Average	**0.094**	**0.100**	**0.0998**	**0.1086**	**0.097**	**0.1034**	**0.1018**	**0.106**	**0.1046**	**0.1076**		

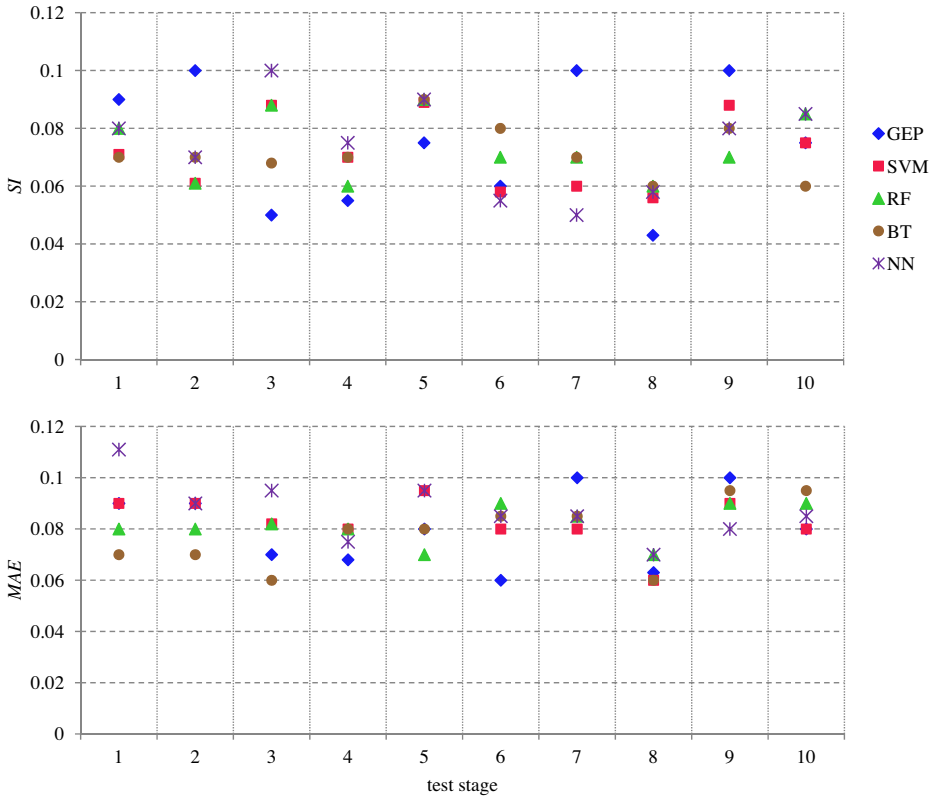

Fig. 6. Split up indicators per test stage of the applied models.

the inputs to output. This can be also observed through analyzing the effect of soil parameters on the BD in Fig. 5. From the figure, the rate of clay, sand, CCE, pH, and OC content effect are 0.0031 (reduction), 0.0445 (increase), 0.0008 (increase), 0.111 (increase), and 0.235 (increase), respectively. Nonlinear effects of these parameters on soil BD were also analyzed (not presented here) and provided similar outcomes. The effect of soil silt on BD (which has been ignored from the input matrix) was very low (lower than 0.001). The main effect is for soil OC, while CCE and soil clay content have the minimum effect, which confirm the results obtained by Al-Qinna and Jaber (2013). However, Al-Qinna and Jaber (2013) stated that silt separate is more effective than the clay on soil BD because clay content in arid soils is low (in non-aggregated form), which cannot be approved for the conditions studied here. Analyzing the soil separates values in Table 1 shows that silt has generally lower values than the both clay and sand separates. Al-Qinna and Jaber (2013) have also suggested a logarithmic relation between the soil OC and BD, which could not be confirmed here, since the linear relation presented the highest effect of OC on BD. Attending the errors of the models, all employed models present the highest negative correlation with soil pH. Soil sand content presents low correlations with the errors of the models.

Further, single-input models were constructed to evaluate the degree of effect of each input variable on models. Table 6 presents the *SI* and *MAE* values of these models. According to the average statistics, the differences between the applied models are marginal. The GEP model has a slightly better accuracy than the other single-input models with respect to average *SI* and *MAE* followed by the SVM model. Comparison with Table 4 clearly shows that the single-input models provide inferior results in estimating soil BD when compared with the quintuple-input models. Average statistics of each variable indicates that the models having CCE and OC as single input respectively perform better than the other models while the pH input general gives the worst estimates.

Split up indices per test stage of the applied quintuple-input models (that use all input parameters) are presented in Fig. 6. The figure clearly displays that the both *SI* and *MAE* indices have notable variations among the test stages for all employed models. GEP models show the highest *SI* and *MAE* variations with $\Delta SI = 0.057$ and $\Delta MAE = 0.041$. Nonetheless, though GEP global performance accuracy is better than the SVM, RF and BT models, it presents lower accuracy (in term of higher error values) in some test stages. According to the *SI* criterion, the GEP model has the best accuracy in the cases 3, 4, 5, 6 and 8 followed by the SVM

which has the lowest *SI* in the cases 2, 6 and 7. However, the BT has the lowest MAE in the cases 1, 2, 3 and 8 followed by the GEP which has the highest accuracy in the cases 4 and 6. The results obtained here demonstrated the need for using the robust k-fold testing for assessing the applied models of soil BD simulations. The results obtained here are belonged to two first soil column depth, so further studies would be necessary to perform similar studies using data from various soil depths.

5. Conclusions

Modeling soil BD using soil parameters through applying the heuristic GEP, NN, RF, SVM and BT models were presented in the current paper. Soil parameters including soil clay, silt, sand, OC, CCE and pH were used to simulate soil BD. Using Gamma test for identifying the best input configuration for feeding the applied models, the soil silt content were omitted from the input matrix and the models were built using the rest parameters. Some previously published pedotrasnfer functions were also used and compared with the heuristic models. A robust k-fold testing was utilized for assessing the applied models. Results showed that the performance accuracies of the heuristic models are generally much better than those of the previously applied pedotransfer functions. Among others, GEP presented the most accurate results in simulating soil BD.

Acknowledgement

This study was partially supported by Department of Soil Science, University of Tehran, Iran. The authors thank the editor and anonymous reviewers for their help in improving the quality of the manuscript.

References

Adams, W.A., 1973. The effect of organic matter on the bulk and true densities of some uncultivated Podzolic soils. Eur. J. Soil Sci. 24, 10–17.
Adrover, M., Farrús, E., Moyà, G., Vadell, J., 2012. Chemical properties and biological activity in soils of Mallorca following twenty years of treated waste water irrigation. J. Environ. Manage. 95, 188–192.
Al-Qinna, M.I., Jaber, S.M., 2013. Predicting soil bulk density using advanced pedotransfer functions in an arid environment. Trans. ASABE 56 (6), 963–976.
Arya, L.M., Paris, J.F., 1981. A physicoempirical model to predict the soil-moisture characteristic from particle-size distribution and bulk-density data. Soil Sci. Soc. Am. J. 45 (6), 1023–1030.
Blake, G.R., Hartge, K.H., 1986. Bulk density. In: Page, A.L. (Ed.), Methods of Soil Analysis, Part 1. American Society of Agronomy, Madison, Wisconsin.
Botula, Y.D., Nemes, A., Ranst, E.V., Mafuka, P., Pue, J.D., Cornelis, W., 2015. Hierarchical pedotransfer functions to predict bulk density of highly weathered soils in Central Africa. Soil Sci. Soc. Am. J. 79 (2), 476–486.
Braun, H.M.H., Kruijne, R., 1994. Soil Conditions. Chapter 3. In: Ritzema, H.P. (Ed.), Drainage Principles and Applications, Publication 27. International Institute for Land Reclamation and Improvement (ILRI), The Netherlands.
Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32.
Curtis, R.O., Post, B.W., 1964. Estimating soil bulk density from organic matter in some Vermont forest soils. Soil. Sci. Soc. Am. Pro. 28, 285–286.
Deschaine, L.M., 2014. Decision Support for Complex Planning Challenges: Combining Expert Systems, Engineering-Oriented Modeling, Machine Learning, Information Theory, and Optimization Technology. Chalmers University of Technology, Sweden, p. 233.
De Vos, B., Van Meirvenne, M., Quataert, P., Deckers, J., Muys, B., 2005. Predictive quality of pedotransfer functions for estimating bulk density of forest soils. Soil Sci. Soc. Am. J. 69, 500–510.
Dexter, A.R., 1988. Advances in Characterization of Soil Structure. Soil Tillage Res. 11 (3–4), 199–238.
Ellert, B.H., Bettany, J.R., 1995. Calculation of organic matter and nutrients stored in soils under contrasting management regimes. Can. J. Soil Sci. 75 (4), 529–538.
Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77, 802–813.
Erdal, H.I., Karakurt, O., 2013. Advancingmonthly streamflow prediction accuracy of CART models using ensemble learning paradigms. J. Hydrol. 477, 119–128.
Evans, D., 2002. The Gamma Test. Data Derived Estimates of Noise for Unknown Smooth Models Using Near Neighbour Asymptotics (Ph.D. thesis). University of Cardiff.
Fausset, L.V. (Ed.), 1994. Fundamentals of Neural Networks: Architectures, Algorithms and Applications. Prentice Hall, Upper Saddle River, NJ.
Federer, C.A., 1983. Nitrogen mineralization and nitrification: depth variation in four New England forest soils. Soil Sci. Soc. Am. J. 47, 1008–1014.
Ferreira, C., 2001. Gene expression programming: a new adaptive algorithm for solving problems. Compl. Syst. 13 (2), 87–129.
Ferreira, C., 2006. Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence. Springer Berlin, Heidelberg, New York, p. 478.
Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: A statistical view of boosting. Ann. Stat. 28 (2), 337–407.
França, S., Cabral, H.N., 2015. Predicting fish species richness in estuaries: Which modelling technique to use? Environ. Modell. Software 66, 17–26.
Francone, F., Deschaine, L.M., 2004. Extending the boundaries of design optimization by integrating fast optimization techniques with machine-code-based, linear genetic programming. Inf. Sci. 161 (3–4), 99–120.
Gee, G.W., Bauder, J.W., 1986. Particle size analysis. In: Klute, A. (Ed.), Methods of Soil Analysis: Part 1. Agronomy Handbook 9. American Society of Agronomy and Soil Science Society of America, Madison, WI, pp. 383–411.
Ghehi, N.G., Nemes, A., Verdoodt, A., Ranst, E.V., Cornelis, W.M., Boeckx, P., 2012. Nonparametric techniques for predicting soil bulk density of tropical rainforest topsoils in Rwanda. Soil Sci. Soc. Am. J. 76 (4), 1172–1183.
Gunn, S.R., 1998. Support Vector Machines for Classification and Regression. In: Technical Report. University of Southampton, England.
Hastie, T., Tibshirani, R., Friedman, J., 2016. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Springer, New York.
Heuvelink, G.B.M., Webster, R., 2001. Modelling soil variation: past, present, and future. Geoderma 100 (3-4), 269–301.
Howell, T.A., Meron, M., 2007. Irrigation scheduling. In: Lamm, F.R., Ayars, J.E., Nakayama, F.S. (Eds.), Microirrigation for Crop Production. Elsevier B.V., pp. 61–130.
Jalabert, S.S.M., Martin, M.P., Renaud, J.P., Boulonne, L., Jolivet, C., Montanarella, L., Arrouays, D., 2010. Estimating forest soil bulk density using boosted regression modelling. Soil Use Manage. 26 (4), 516–528.
Jeffrey, D.W., 1970. A note on the use of ignition loss as a means for the approximate estimation of soil bulk density. J. Ecol. 58, 297–299.
Jones, A.J., Evans, D., Margetts, S., Durrant, P.J., 2002. Heuristic and Optimization for Knowledge Discovery. Idea Group Publishing, Hershey, PA. Chapter IX.
Kisi, O., Dailr, A.H., Cimen, M., Shiri, J., 2012. Suspended sediment modeling using genetic programming and soft computing techniques. J. Hydrol. 450–451, 48–58.
Kisi, O., Shiri, J., 2012. River suspended sediment estimation by climatic variables implication: Comparative study among soft computing techniques. Comput. Geosci. 43, 73–82.
Kisi, O., Shiri, J., Tombul, M., 2013. Modeling rainfall-runoff process using soft computing techniques. Comput. Geosci. 51, 108–117.
Koza, J.R., 1992. Genetic Programming: On the Programming of Computers by Means of Natural Selection. The MIT Press, Cambridge, MA, p. 840.
Lal, R., Kimble, J.M., 2001. Importance of Soil Bulk Density and Methods of Its Importance. In: Lal, R., Kimble, J.M., Follett, R.F., Stewart, B.A. (Eds.), Assessment Methods for Soil Carbon. Lewis Publishers, London, pp. 31–44.
Lampurlanes, J., Cantero-Martinez, C., 2003. Soil bulk density and penetration resistance under different tillage and crop management systems and their relationship with barley root growth. Agron. J. 95 (3), 526–536.
Manrique, L.A., Jones, C.A., 1991. Bulk densities of soils in relation to soil physical and chemical properties. Soil Sci. Soc. Am. J. 55, 476–481.
Marti, P., Shiri, J., Duran-Ros, M., Arbat, G., Cartagena, F.R., Puig-Bargues, J., 2013. Artificial neural networks vs. gene expressions programming for estimating outlet dissolved oxygen in micro irrigation sand filters fed with effluents. Comput. Electron. Agric. 99, 176–185.
Mermoud, A., Xu, D., 2006. Comparative analysis of three methods to generate soil hydraulic functions. Soil Tillage Res. 87, 89–100.
Minasny, B., McBratney, A.B., 2002. The neuro-m methods for fitting neural network parametric pedotransfer functions. Soil Sci. Soc. Am. J. 66, 352–361.
Nelson, R.E., 1982. Carbonate and gypsum. In: Page, A.L. (Ed.), Methods of Soil Analysis: Part 1. Agronomy Handbook 9. American Society of Agronomy and Soil Science Society of America, Madison, WI, pp. 181–197.
Nelson, D.W., Sommers, L.P., 1986. Total carbon, organic carbon and organic matter. In: Page, A.L. (Ed.), Methods of Soil Analysis: Part 2. Agronomy Handbook 9. American Society of Agronomy and Soil Science Society of America, Madison, WI, pp. 539–579.
Patil, N.G., Chaturvedi, A., 2012. Estimation of bulk density of waterlogged soils from basic properties. Arch. Agron. Soil Sci. 58 (5), 499–509.
Pfleger K., John G., and Kohavi R., 1994. Irrelevant features and the subset selection problem, Machine Learning. In: Proceedings of the Eleventh International Conference, pp. 121–129.
Phuong, T.M., Lin, Z., Altman, R.B., 2005. Choosing SNPs using feature selection. In: Proceedings/IEEE Computational Systems Bioinformatics Conference, CSB. IEEE Computational Systems Bioinformatics Conference, pp. 301–309.
Rice, J.S., Emanuel, R.E., Vose, J.M., Nelson, S.A.C., 2015. Continental U.S. streamflow trends from, to 2009 and their relationships with watershed spatial characteristics. Water Resour. Res. 51. http://dx.doi.org/10.1002/2014WR016367.

Rice, J.S., Emanuel, R.E., Vose, J.M., 2016. The influence of watershed characteristics on spatial patterns of trends in annual scale streamflow variability in the continental U.S. J. Hydrol. 540, 850–860.
Rodríguez-Lado, L., Rial, M., Taboada, T., Martínez Cortizas, A., 2015. A pedotransfer function to map soil bulk density from limited data. Procedia Environ. Sci. 27, 45–48.
Rodrigues, M., Riva, J., 2014. An insight into machine-learning algorithms to model human-caused wildfire occurrence. Environ. Modell. Software 57, 192–201.
Schaap, M.G., Leij, F.J., 1998. Using neural networks to predict soil water retention and soil hydraulic conductivity. Soil Tillage Res. 47, 37–42.
Schoeneberger, P.J., Wysocki, D.A., Benham, E.C., Broderson, W.D., 2002. Field Book for Describing and Sampling Soils, Version 2.0. NRCS-National Soil Survey Center, Lincoln, NE.
Shiri, J., Kisi, O., 2012. Estimation of daily suspended sediment load by using wavelet conjunction models. Asce. J. Hydrol. Eng. 17 (9), 986–1000.
Shiri, J., Kisi, O., Landeras, G., Lopez, J.J., Nazemi, A.H., Stuyt, L.C.P.M., 2012. Daily reference evapotranspiration modeling by using genetic programming approach in the Basque Country (Northern Spain). J. Hydrol. 414–415, 302–316.
Shiri, J., Marti, P., Singh, V.P., 2014a. Evaluation of gene expression programming approaches for estimating daily evaporation through spatial and temporal data scanning. Hydrol. Process. 28 (3), 1215–1225.
Shiri, J., Nazemi, A.H., Sadraddini, A.A., Landeras, G., Kisi, O., Fakheri Fard, A., Marti, P., 2014b. Comparison of heuristic and empirical approaches for estimating reference evapotranspiration from limited inputs in Iran. Comput. Electron. Agric. 108, 230–241.
Shiri, J., Keshavarzi, A., Kisi, O., Iturraran-Viveros, U., Bagherzadeh, A., Mousavi, R., Karimi, S., 2017. Modeling soil cation exchange capacity using soil parameters: Assessing the heuristic models. Comput. Electron. Agric. 135, 242–251.
Thomas, G.W., 1996. Soil pH and soil acidity. In: Page, A.L. (Ed.), Methods of Soil Analysis: Part 2. Agronomy Handbook 9. American Society of Agronomy and Soil Science Society of America, Madison, WI, pp. 475–490.
Tisseuil, C., Vrac, M., Lek, S., Wade, A.J., 2010. Statistical downscaling of river flows. J. Hydrol. 385, 279–291.
Toming, K., Kutser, T., Tuvikene, L., Viik, M., Noges, T., 2016. Dissolved organic carbon and its potential predictors in eutrophic lakes. Water Res. 102, 32–40.
Tsui, P.M., Jones, A.J., Oliveira, A.G., 2002. The construction of smooth models using irregular embeddings determined by a Gamma test analysis. Neural Comput. Appl. 10, 318–329.
Vapnik, V., Golwich, S., Smola, A.J., 1997. Support vector method for function approximation, regression estimation, and signal processing. In: Mozer, M., Jordan, M., Petsche, T. (Eds.), Advances in Neural Information Processing Systems 9. Neural Information Processing Systems Foundation, Inc, pp. 281–287.
Xiangsheng, Y., Guosheng, L., Yanyu, Y., 2016. Pedotransfer functions for estimating soil bulk density: a case study in the Three-River Headwater region of Qinghai Province, China. Pedosphere 26 (3), 362–373.