Rastgou, M., Bayat, H., Mansoorizadeh, M. and Gregory, A. S. 2020. Estimating the soil water retention curve - comparison of multiple nonlinear regression approach and random forest data mining technique. *Computers and Electronics in Agriculture.* 174, p. 105502.

Manuscript Details

Manuscript number	COMPAG_2019_2240_R2
Title	Estimating the soil water retention curve: comparison of multiple nonlinear regression approach and random forest data mining technique
Article type	Research Paper

Abstract

This study evaluates the performance of the random forest (RF) method on the prediction of the soil water retention curve (SWRC) and compares its performance with those of nonlinear regression (NLR) and Rosetta-based pedotransfer functions (PTFs), which has not been reported so far. Fifteen RF and NLR-based PTFs were constructed using readily-available soil properties for 223 soil samples from Iran. The general performance of RF and NLR-based PTFs was quantified by the integral root mean square error (IRMSE), Akaike's information criterion (AIC) and coefficient of determination (R2). The results showed that the accuracy of the RF-based PTFs was significantly (P<0.05) better than the NLR-based PTFs, and that the reliability of the NLR-based PTFs was significantly (P<0.01) better than the RF-based PTFs and all of the Rosetta-based PTFs. The average values of the IRMSE, AIC and R2 of the RF method were 0.041 cm3 cm-3, -16997.7, and 0.987, and 0.053 cm3 cm-3, -15547.5, and 0.981 for the training and testing steps of all PTFs, respectively, whereas the values for the NLR method were 0.046 cm3 cm-3, -16616.4, and 0.984, and 0.048 cm3 cm-3, -16355.6, and 0.983 for the training and testing steps, respectively. The PTF5 of the RF and NLR methods, with inputs of sand and clay contents, bulk density, and the water content at field capacity and permanent wilting point, had the greatest R2 values (0.987 and 0.989, respectively), and the lowest IRMSE values (0.039 and 0.032 cm3 cm-3, respectively) compared to other PTFs for the testing step. Overall, the RF method had less reliability for the prediction of the SWRC compared to the NLR method due to overprediction, uncertainty of determination of forest scale and instability in the testing step. These findings could provide the scientific basis for further research on the RF method.

Keywords	pedotransfer functions; soil water retention curve; soil texture; soil structure; van Genuchten.

Corresponding Author	Hossein Bayat
Corresponding Author's Institution	Bu Ali Sina University
Order of Authors	Mostafa Rastgou, Hossein Bayat, Muharram Mansoorizadeh, Andrew Gregory

Suggested reviewers	Jorge Werneck Lima, Estela. N. Hepper, Sabit Ersahin, Masoud Davari, Hojat Emami

Submission Files Included in this PDF

File Name [File Type]

Cover letter.docx [Cover Letter]

Answers to the comments.docx [Response to Reviewers]

Revision changes marked-pic.docx [Review Reports]

Highlights.docx [Highlights]

Clean File.docx [Manuscript File]

conflict of interests.docx [Conflict of Interest]

Author statement.docx [Author Statement]

To view all the submission files, including those not included in the PDF, click on the manuscript title on your EVISE Homepage, then click 'Download zip file'.

Dear Prof. He,

Thanks a lot for your helpful advice and the reviewers' useful comments and suggestions on our manuscript. We modified and revised the manuscript accordingly and details of the corrections are described in the “Answer to the comments” file, point by point. One of the co-authors is a native English speaker and he has thoroughly checked and corrected spelling and grammatical errors. Then two versions of the manuscript were resubmitted to your journal: a version of the revised manuscript showing the new/changed text using track changes and a clean version of the revised manuscript. It would be appreciated if you could please kindly let me know if there is any other deficiency with our manuscript. We look forward to your positive response.

With best regards

Hossein Bayat

Hossein Bayat: Department of Soil Science, Faculty of Agriculture, Bu Ali Sina University, Hamedan, Iran. Postal Address: Department of Soil Science, Faculty of Agriculture, Bu Ali Sina University, Hamedan, Iran.

E-mail; h.bayat@basu.ac.ir. Tel: +98-918-8188378 and +98-81-34424189

Dear Prof. He,

Thanks a lot for your helpful advice and the reviewers' useful comments and suggestions on our manuscript. We modified and revised the manuscript accordingly and details of the corrections are described below point by point. One of the co-authors is a native English speaker and he has thoroughly checked and corrected spelling and grammatical errors. Then two versions of the manuscript were resubmitted to your journal: a version of the revised manuscript showing the new/changed text using track changes and a clean version of the revised manuscript. It would be appreciated if you could please kindly let me know if there is any other deficiency with our manuscript. We look forward to your positive response.

With best regards

Hossein Bayat

Note: Page numbers and line numbers that are given in this file are according to those of the "Revision, changes marked" file.

Comments from the editors and reviewers:

-Reviewer 1

-The authors replied to only partially to my comments. In fact, they replied to those comments for Author but not to those for Editor. Did they not receive the comments for Editor? My decision is still Major revision since many points raised before did not receive answers. I put them again below. **Page and line numbers refer to the original version of the manuscript, not the revised one.**

Ans**:**

We apologize for the inconvenience, but unfortunately we did not receive the comments in the first round of Review. Now, we have modified and revised the manuscript according to your comments and details of the corrections are described below point by point. The authors are grateful for your comments in improving the content and structure of the manuscript.

First of all, the use of random forest to PTF is not completely new as may be deduced from the manuscript (page 5, lines 88 and 89); in contrast, there are published papers that dealt with random forest like Toth et al (2014), Araya et al (2019), Gunarathna et al (2019), and Szabo et al (2019). Also, the authors gave few examples of the use of statistical and data mining techniques but only one example for the nearest neighbor (page 4, line 75) as if it is the only published work whereas there are many other examples like Botula et al (2013), Haghverdi et al (2015), Nguyen et al (2017), Gunarathna et al (2019), etc.

Ans:

Thank you so much. A review of literatures (Toth et al. (2014), Araya et al. (2019), Gunarathna et al. (2019), and Szabo et al. (2019)) revealed that the RF data mining technique has only been applied to predict point-based PTFs of the SWRC including field capacity and permanent wilting point or saturated hydraulic conductivity, but it has not been used for developing parametric-based PTFs of the van Genuchten model parameters, so far. Finally, the review of literatures has been modified completely as follows:

"So far, few studies have been carried out on the application of the RF method in soil science. Tóth et al. (2014) applied the RF method to analyze the relationship between soil water content at four matric suctions (0.1, 33, and 1500 kPa, and 150 MPa) and Hungarian soil map information. They found that the importance of soil properties in the prediction of the soil water content varied according to soil type and matric suction. Recently Szabó et al. (2019) have developed PTFs based on RF and geostatistics methods to map soil hydraulic properties, such as water contents at saturation, field capacity and wilting point, for the Balaton catchment area in Hungary. Araya and Ghezzehei (2019) compared the performances of four machine-learning algorithms including the k-nearest neighbors (kNNs), support vector regression (SVR), RF, and boosted regression tree (BRT) for prediction of saturated hydraulic conductivity. They found that the BRT model outperformed the other algorithms closely followed by the RF model. Gunarathna et al. (2019a) tested three machine-learning algorithms including ANN, kNN, and RF to estimate volumetric water content at matric suctions of 10, 33 and 1500 kPa for soils in Sri Lanka. They recommended that the PTFs to be developed using the RF algorithm. Ließ et al. (2012) studied uncertainty in the spatial prediction of soil texture by comparison of the RF and regression tree techniques for 56 soil profiles and found that the former method provided a better result. Also, Wiesmeier et al. (2011) utilized the RF technique to develop digital mapping of the soil organic matter content in 120 soil profiles. They found that the prediction accuracy of the RF modeling was acceptable. A review of literatures therefore revealed that the RF data mining technique has been applied to develop PTFs to predict specific points of the SWRC, such as field capacity and permanent wilting point, or particular properties such as saturated hydraulic conductivity, but it has not been used to develop parametric-based PTFs of the van Genuchten model parameters, so far (Pages 5-6, lines 84-109).

Also, we have added new literatures, in which statistical and data mining techniques have been used, to the introduction section of the manuscript, as follows:

Botula et al. (2013): Page 4, line 77.

Haghverdi et al. (2015): Page 4, line 77.

Nguyen et al. (2017): Page 4, line 78.

Gunarathna et al. (2019a): Page 4, line 74.

Gunarathna et al. (2019a): Page 4, line 77.

Gunarathna et al. (2019b): Page 4, line 72.

Khlosi et al. (2016): Page 4, line 78.

At page 6, lines 115-122 (section 2.2.), authors are presenting results in the Material and Methods section. Therefore, this section should be moved to Results and Discussion section. By the way the maximum clay content is 48% (Table 1), so the sentence should be rewritten accordingly.

Ans:

Following your suggestion, section *2.2* was moved and is now section *3.1* in the ”Results and discussion” section (Page 15, lines 283-293). Also, the sentence has been rewritten as follows: "It can be seen that the average and maximum of clay content were 21.4 and 48%, respectively" (Page 15, lines 285-286).

The same remark as above applies to page 7, lines 135-144 (section 2.4.): to move to Results and Discussion section.

Ans:

Following your suggestion, section *2.4* was moved and is now section *3.2* in the ”Results and discussion” section (Pages 15-16, lines 294-331).

In addition, at line 138, authors are listing the soil properties that have high correlation with van Genuchten parameters. They did not mention thetapPWP even it had high correlation coefficients!

Ans:

Thank you so much. It is a good point. This point has been mentioned in the manuscript.

Therefore, the sentence has been modified as follows:

"Clay and sand contents, θ_{FC}, θ_{PWP}, d_g and OM had the greatest significant correlations with the parameters of the van Genuchten model (Fig. 4), which are consistent with other studies (Dexter et al., 2008; Nemes et al., 2006). For example, the correlation coefficients between clay content and θ_s ($r = 0.323$) is close to that between the OM and θ_s ($r = 0.268$). Also, the results showed that there were significant correlations between θ_{PWP} and input variables of clay content (+), sand content (–), BD (–), OM (+) and K_s (–), and also between θ_{PWP} and θ_s (+) and n (–) parameters of the van Genuchten model (Fig. 4) (Fig. 4). Botula et al. (2012) also found the same observation for the correlation of θ_{PWP} with sand and clay contents and BD of tropical Lower Congo soils (Page 16, lines 299-307).

Also, at lines 143 and 144, the authors stated that there was no correlation between van Genuchten parameters and Ks whereas they used this soil property in PTF14 and PTF15. Could they explain why they considered Ks even if it not correlated to van Genuchten parameters?

Ans:

There are many cases, where two variables might not show a strong simple correlation, but may show a strong association in the regression, along with other predictors. In other words, the simple correlation coefficient is a way to show the relationship between independent and dependent variables, but it cannot show a model for the relationship between these two variables, when other independent variables have been used in a multiple regression (Simmons et al., 2011). The result of multiple regression analysis with backward selection method showed that the K_s variable remained in the PTF14 and PTF15 for all the van Genuchten model parameters. Some of the regression equations with backward selection method are shown in the following as examples:

θ_r=-0.69+0.22×Clay+0.278×Sand+0.20×K_s, R=0.31**

α=-3.72+0.23×Clay+0.17×BD+0.282×K_s, R=0.33** and

n=-1.76+0.24×Sand+0.164×K_s, R=0.30**.

On the other hand, the non-linear correlations between variables are very important in this study. Both the multiple NLR approach and RF data mining technique are non-linear prediction methods. Fig. 4 only shows simple linear correlation between variables, but there may be non-linear correlations between variables, which may affect the estimation of the dependent variables. For example, the results of non-linear correlations showed that K_s had strong correlations with θ_s and α of the van Genuchten model parameters by logarithmic (θ_s=0.652-0.027×lnK_s, R=0.62**) and power (α=0.007×$K_s^{0.283}$, R=0.57**) equations, respectively, which were greater than their simple correlations (Pages 16-17, lines 310-328). In support of this claim, the results showed that by adding OM and/or K_s as predictors in the PTFs 13, 14 and 15, the accuracy (Fig. 5B) and reliability (Fig. 6B) of the prediction of the SWRC improved by 16, 13, 17 and 7.1, 6.3, 6.9%, respectively, compared to the PTF3 in terms of the *IRMSE* criterion in the RF method (Pages 25-26, lines 517-520).

At page 8, line 152, the authors assessed multicollinearity using the variance inflation factor (VIF) in the Material and Methods section but they reported nothing about this in the Results and

Discussion section; although they mentioned that silt content was eliminated to avoid multicollinearity (line 164).

Ans:

You are completely right. The values of variance inflation factor (*VIF*) for all PTFs have been added to the manuscript. Therefore, the text has been modified as follows:

"Before developing PTFs, all variables were evaluated by Kolmogorov-Smirnov normality and multicollinearity tests by the SPSS 24 software (IBM, 2016). The degree of multicollinearity in the PTFs was tested by the variance inflation factor (VIF=1/1-R^2_j, where R^2_j is the R^2 value obtained by regressing the j^{th} predictor on the remaining predictors) (Hocking, 2013). Also, to avoid multicollinearity between textural contents, the silt fraction was not used as a predictor" (Page 9, lines 157-161). Results of the multicollinearity analysis (*VIF)* are shown in Table 3. The *VIF* values showed low levels of multicollinearity among the independent variables (VIF<10) (Khodaverdiloo et al., 2011) (Page 17, lines 334-336).

Table 3- The variance inflation factor (*VIF*) values for normalized form of input variables.

PTFs	Clay* (%)	Sand (%)	BD$^{\$}$ (g cm^{-3})	θ_{FC} (cm^3 cm^{-3})	θ_{PWP} (cm^3 cm^{-3})	d_g (mm)	δ_g (-)	TP (cm^3 cm^{-3})	OM (%)	K_s (cm day^{-1})
PTF2	1.42	1.42								
PTF3	1.43	1.52	1.10							
PTF4	1.45	1.56	1.25	1.31						
PTF5	1.79	1.58	1.27	2.48	2.56					
PTF6						1.00	1.00			
PTF7			1.11			1.11	1.01			
PTF8			1.25	1.33		1.01	1.22			
PTF9			1.28	2.50	2.73	1.34	1.22			

PTFs	Clay* (%)	Sand (%)	BD$ (g cm^{-3})	θ_{FC} (cm^3 cm^{-3})	θ_{PWP} (cm^3 cm^{-3})	d_g (mm)	δ_g (-)	TP (cm^3 cm^{-3})	OM (%)	K_s (cm day^{-1})
PTF10	1.55	1.43						1.11		
PTF11	1.58	1.46		1.32				1.26		
PTF12	1.60	1.79		2.49	2.56			1.28		
PTF13	1.48	1.65	1.25						1.14	
PTF14	1.55	1.64	1.14							1.06
PTF15	1.55	1.65	1.25						1.15	1.06

* Normalized form of input variables is available in Table 2.

$. A list of abbreviations is available in the notation box.

Page 10, lines 198-203, the authors used 20-fold cross validation: the question why the authors used this specific k value and not, for example, 10 which is the most commonly used one in cross validation?

Ans:

In the present study, the k-fold cross validation approach (Efron and Tibshirani, 1994) was utilized to obtain training and testing datasets for each PTF. The number of folds (i. e., k) was obtained by trial and error. To do so, some PTFs, selected randomly, were developed with 10, 15 and 20-fold cross-validation. Then, the k value which resulted in the best performance of the PTFs, was selected to develop all PTFs in this study. The results showed that 20-fold cross validation performed better than the other folds in most of the PTFs (Table 1). Therefore, 20-fold cross validation was selected to develop PTFs in this study (page 11, lines 201-207).

Table 1- The results of 10, 15 and 20-fold cross-validation (k) for van Genuchten model parameters of the soil water retention curve derived from nonlinear regression (NLR) and

random forest (RF) techniques based on root mean square error (*RMSE*) for pedotransfer functions PTF 3, 5 and 11 in the train and test datasets.

			θ_r			θ_s			α			n		
			RMSE			*RMSE*			*RMSE*			*RMSE*		
			Train	Test	Mean	Train	Test	Mean	Train	Test	Mean	Train	Test	Mean
PTF3	k=10	NLR	0.058	0.060	0.059	0.063	0.065	0.064	1.017	1.037	1.027	0.426	0.436	0.431
		RF	0.052	0.061	0.056	0.058	0.073	0.066	0.893	1.084	0.989	0.374	0.442	0.408
	k=15	NLR	0.058	0.060	0.059	0.064	0.064	0.064	1.017	1.030	1.024	0.426	0.434	0.430
		RF	0.052	0.061	0.057	0.058	0.070	0.064	0.894	1.033	0.964	0.374	0.441	0.408
	k=20	NLR	0.058	0.060	0.059	0.064	0.064	0.064	0.064	0.064	0.064	0.426	0.437	0.432
		RF	0.051	0.060	0.056	0.057	0.071	0.064	0.057	0.071	0.064	0.368	0.442	0.405
PTF5	k=10	NLR	0.051	0.053	0.052	0.053	0.054	0.054	0.764	0.796	0.780	0.380	0.397	0.389
		RF	0.043	0.056	0.050	0.046	0.056	0.051	0.675	0.869	0.772	0.327	0.411	0.369
	k=15	NLR	0.051	0.053	0.052	0.053	0.055	0.054	0.764	0.790	0.777	0.381	0.399	0.390
		RF	0.044	0.054	0.049	0.046	0.055	0.050	0.679	0.848	0.763	0.329	0.421	0.375
	k=20	NLR	0.051	0.053	0.052	0.053	0.055	0.054	0.765	0.789	0.777	0.381	0.399	0.390
		RF	0.042	0.054	0.048	0.044	0.054	0.049	0.654	0.842	0.748	0.316	0.412	0.364
PTF11	k=10	NLR	0.058	0.061	0.060	0.065	0.067	0.066	1.018	1.052	1.035	0.431	0.448	0.440
		RF	0.050	0.061	0.056	0.047	0.057	0.052	0.770	0.978	0.874	0.370	0.443	0.406
	k=15	NLR	0.058	0.061	0.060	0.065	0.067	0.066	1.019	1.037	1.028	0.432	0.447	0.439
		RF	0.050	0.060	0.055	0.047	0.057	0.052	0.770	1.009	0.889	0.369	0.450	0.410
	k=20	NLR	0.058	0.060	0.059	0.065	0.065	0.065	1.020	1.024	1.022	0.432	0.439	0.435
		RF	0.049	0.061	0.055	0.046	0.056	0.051	0.745	0.964	0.855	0.361	0.443	0.402

Also, the authors used data from 6 different provinces and 2 soil depths. I wonder if they took into consideration these two distinguishing factors when they split their data during k-fold cross validation into training and testing subsets.

Ans:

All soil samples, which have been collected from 6 different provinces and 2 soil depths, have been assumed as one database and training and testing data have been selected randomly from the database (which have been included all soil samples). In other words, we have not taken into consideration these two distinguishing factors (province or depth of sampling) when we split all data during k-fold cross validation into training and testing subsets.

Page 10, line 208 and equation (2): the authors noted the number of input variables by n; there may be confusion with the fourth parameter of van Genuchten model (page 7, equation (1) and line 131)! Here n may be replaced by p (the number of input variables like in the AIC definition at page 13, equation (6). By the way the authors should use the same letter: p and not P (line 256)!

Ans:

Thank you so much. The required correction has been done (Page 11, line 216).

Page 13, lines 258-260: the average values can be compared using the analysis of variance (ANOVA) method and, once they are significantly different, we can use some posthoc tests like the Duncan test. However, it is not clear what was compared: all the 15 PTFs for both RF and NLR, and even from Rosetta for the testing datasets (Figures 6 and 7, graphs B) or the 2 or 3 methods (NLR, RF, and Rosetta) separately for each of the 15 PTFs (page 14, lines 270-273). If it is the former case, Duncan test is useless since it compares 30 mean values (and even 35 if we consider Rosetta in addition to NLR and RF) and consequently some PTFs are belonging to 2 or 3 different groups (like PTF4 RF, PTF5 NLR, etc. with abc letters) for training data sets (Figure 6) and even more for the testing dataset (4 letters like h-k or i-l on Figure 7). Moreover, this statistical comparison was done only for IRMSE but not for the 3 other cross validation criteria (IME, R^2, and AIC). Is there any explanation?

Ans:

Due to the fact that the performance of both methods was evaluated for all samples, therefore the mean comparison test can be used to compare the predicting accuracy and reliability of the RF and NLR methods. In other words, to determine whether the differences in the accuracy and reliability of the RF and NLR methods are random or real, the mean comparison test could be performed. One of the aims of the present study was to investigate the performance of each method in different

PTFs. In other words, the performance of each method in each PTF was important to the users. "To evaluate the performance of each method in different PTFs, the effect of method as the first factor at two levels in the training step (*i.e.*, NLR and RF methods) and at three levels in the testing step (*i.e.*, NLR, RF and Rosetta methods), and the different PTFs as the second factor at 15 levels (PTF1 to PTF15), were investigated using a two-way analysis of variance (ANOVA) with a randomized complete block design, based on the *IRMSE* of prediction of the SWRC" (Pages 14-15, lines 270-275). Table 4 shows the results of the ANOVA of the *IRMSE* of prediction of the SWRC by different methods and PTFs. The effect of methods and PTFs, and their interaction, on the *IRMSE* was significant at P<0.01, 0.01 and 0.05, respectively, in the training step, and at P<0.01, 0.01 and 0.01, respectively, in the testing step. Therefore, we focus on the results and discussion of the comparison of the method × PTF interaction effects (Page 18, lines 340-346).

The *IRMSE* criterion calculates the total error, including bias and random errors, and is a more appropriate criterion for evaluating the accuracy and reliability of the RF and NLR methods compared to other criteria (Chai and Draxler, 2014). Therefore, to compare the predicting accuracy and reliability of the RF and NLR methods, the average values of the *IRMSE* was compared with Duncan's test by MathWorks (2018) software (Page 15, lines 275-280).

Table 4- Analysis of variance of the integral root mean square error (*IRMSE*) of the prediction of the soil water retention curve by different methods (nonlinear regression and random forest) and pedotransfer functions (PTFs 1-15) for both the train and test datasets.

	Source	Degree freedom	Mean square	*F*-value	*P*-value
Train	Repeat (Block)	222	0.007	19.09	<0.0001
	PTFs	14	0.062	180.68	<0.0001
	Methods	1	0.038	109.69	<0.0001
	PTFs × Methods	14	0.001	1.78	0.0356
	Error	6288	0.0003		
Test	Repeat (Block)	222	0.010	16.04	<0.0001
	PTFs	14	0.073	117.22	<0.0001
	Methods	2	0.656	1056.43	<0.0001

PTFs × Methods	18	0.002	3.68	<0.0001
Error	7398	0.0006		

At page 19, lines 385-387: the authors are discussing the correlation between thetar and referring to Figure 2 whereas correlation coefficients between thetar and soil proprieties were not included in this figure!

Ans:

The correlation test was not performed for the θ_r variable, because its value was zero in 138 out of 223 soil samples, as has been reported in other studies (Campbell and Horton Jr, 2002; Rawls et al., 1991; Tomasella et al., 2000) for θ_r variable (Pages 15-16, lines 296-299). Therefore, the sentence has been rewritten as follows: "Therefore, input variables of the textural contents or statistics can influence the residual saturation region of the SWRC. However, soil water content at the dry end (high matric suctions) of the SWRC is primarily determined by textural contents (Hillel, 1998) " (Pages 23-24, lines 470-473).

In addition, the whole subsection 3.1.2.2. is about the importance of the introduction of Ks into PTF 14 and 15 whereas there was no correlation between van Genuchten parameters and Ks. How the authors can explain the added value of Ks to the last 2 PTFs even in the absence of significant correlation?

Ans:

It has been answered in pages 5-6, lines 113-135 of this file.

Furthermore, at page 21, lines 442 and 443, the authors said that Ks is correlated to soil texture and TP variables whereas it is correlated only to clay content (Figure 2) but not to sand nor to TP.

Ans:

Thank you so much. The sentence has been rewritten as follows:

"The correlation results showed (Fig. 4) that K_s can be strongly influenced by clay content and textural statistics (d_g and δ_g)" (Page 26, lines 524-525).

-Reviewer 2

-I thank the authors for their through addressing my queries and completing the recommended revisions. The authors should address following points.

Ans:

Thank you so much. Your comments helped us a lot to improve the manuscript.

1. Revise L45-46 as follows: “These findings could provide the scientific basis for further research on the RF method.”

Ans:

It has been done (page 2, lines 45-46).

2. I could not find the following revision in the revised manuscript, please recheck for its existence.

L104-105: What do you mean by “topsoil” and “subsoil”? Do you mean A and B horizons or tillage depth? Be specific. Also, what do you mean with layer in “depending on thickness of layers”?

Ans:

”topsoil” and ”subsoil” refer to A and B horizons, respectively. It was corrected in the manuscript. Since the sampling was done from different locations of the various provinces, the topsoil and subsoil layers of soil at different locations had different depths and thicknesses, and samples were taken from the center of each layer. Therefore, the samples were taken from different depths, depending on thickness of the A and B layers.

Ans:

Thank you so much. It has been arranged as follows:

Since the sampling was done from different locations of the various provinces, the topsoil and subsoil layers of soil at different locations had different depths and thicknesses. We collected samples from the center of the topsoil and subsoil layers, which represented the pedological A and B horizons, respectively. The sampling depths varied from 10 to 35 cm for topsoil (208 samples) and from 20 to 45 cm for subsoil (15 samples), reflecting the variation in the soil profiles (pages 6-7, lines 123-130).

3. I do recommend the authors go over the manuscript for mistakes of grammar, typos, sentence structure, and so on before sending their final copy to the editor.

Ans:

We thank the reviewer for this point. The co-author for whom English is their first language has been through the manuscript thoroughly and corrected all errors in spelling and grammar.

Eventually;
As it was described point by point, the manuscript was revised significantly.

Acknowledgements

The authors are deeply grateful to anonymous reviewers and the editor for their helpful comments on the manuscript.

References

Araya, S.N., Ghezzehei, T.A., 2019. Using Machine Learning for Prediction of Saturated Hydraulic Conductivity and Its Sensitivity to Soil Structural Perturbations. Water Resour. Res. 55(7), 5715-5737.

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

Botula, Y.-D., Cornelis, W.M., Baert, G., Mafuka, P., Van Ranst, E., 2013. Particle size distribution models for soils of the humid tropics. Journal of Soils and Sediments 13(4), 686-698.

Campbell, G.S., Horton Jr, R., 2002. Methods of Soil Analysis: Part 4, Physical Methods. Soil Sci. Soc. Am.

Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)?– Arguments against avoiding RMSE in the literature. Geosci. Model Dev. 7(3), 1247-1250.

Dexter, A., Czyż, E., Richard, G., Reszkowska, A., 2008. A user-friendly water retention function that takes account of the textural and structural pore spaces in soil. Geoderma 143(3), 243-253.

Efron, B., Tibshirani, R.J., 1994. An introduction to the bootstrap. CRC press.

Gunarathna, M., Sakai, K., Nakandakari, T., Momii, K., Kumari, M., 2019a. Machine Learning Approaches to Develop Pedotransfer Functions for Tropical Sri Lankan Soils. Water 11(9), 1940.

Gunarathna, M., Sakai, K., Nakandakari, T., Momii, K., Kumari, M., Amarasekara, M., 2019b. Pedotransfer functions to estimate hydraulic properties of tropical Sri Lankan soils. Soil Till. Res. 190, 109-119.

Haghverdi, A., Leib, B.G., Cornelis, W.M., 2015. A simple nearest-neighbor technique to predict the soil water retention curve. Transactions of the ASABE 58(3), 697-705.

Hillel, D., 1998. Environmental soil physics: Fundamentals, applications, and environmental considerations. Academic press.

Hocking, R.R., 2013. Methods and applications of linear models: regression and the analysis of variance. John Wiley & Sons.

IBM, C., 2016. IBM SPSS Statistics for Windows, Version 24.0. Armonk, NY: IBM Corp.

Khlosi, M., Alhamdoosh, M., Douaik, A., Gabriels, D., Cornelis, W., 2016. Enhanced pedotransfer functions with support vector machines to predict water retention of calcareous soil. Eur. J. Soil Sci. 67(3), 276-284.

Khodaverdiloo, H., Homaee, M., van Genuchten, M.T., Dashtaki, S.G., 2011. Deriving and validating pedotransfer functions for some calcareous soils. J. Hydrol. 399(1), 93-99.

Ließ, M., Glaser, B., Huwe, B., 2012. Uncertainty in the spatial prediction of soil texture: comparison of regression tree and Random Forest models. Geoderma 170, 70-79.

MathWorks, 2018. MATLAB: the language of technical computing. Inc., Natick, Massachusetts, United States.

Nemes, A., Rawls, W.J., Pachepsky, Y.A., 2006. Use of the nonparametric nearest neighbor approach to estimate soil hydraulic properties. Soil Sci. Soc. Am. J. 70(2), 327-336.

Nguyen, P.M., Haghverdi, A., De Pue, J., Botula, Y.-D., Le, K.V., Waegeman, W., Cornelis, W.M., 2017. Comparison of statistical regression and data-mining techniques in estimating soil water retention of tropical delta soils. Biosyst. Eng. 153, 12-27.

Rawls, W., Gish, T., Brakensiek, D., 1991. Estimating soil water retention from soil physical properties and characteristics, Advances in soil science. Springer, pp. 213-234.

Simmons, J.P., Nelson, L.D., Simonsohn, U., 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol. Sci. 22(11), 1359-1366.

Szabó, B., Szatmári, G., Takács, K., Laborczi, A., Makó, A., Rajkai, K., Pásztor, L., 2019. Mapping soil hydraulic properties using random forest based pedotransfer functions and geostatistics. Hydrol. Earth Syst. Sci. 23(6), 2615-2635.

Tomasella, J., Hodnett, M.G., Rossato, L., 2000. Pedotransfer functions for the estimation of soil water retention in Brazilian soils. Soil Sci. Soc. Am. J. 64, 327-338.

Tóth, B., Makó, A., Toth, G., 2014. Role of soil properties in water retention characteristics of main Hungarian soil types. J. Cent. Eur. Agric. 15(2), 137-153.

Wiesmeier, M., Barthold, F., Blank, B., Kögel-Knabner, I., 2011. Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. Plant Soil 340(1), 7-24.

Estimating the soil water retention curve: comparison of multiple nonlinear regression approach and random forest data mining technique

M. Rastgou[1], H. Bayat[2]*, and M. Mansoorizadeh[3], Andrew S. Gregory[4]

[1] **Mostafa Rastgou:** Ph. D. Student of Soil Science, Department of Soil Science, Faculty of Agriculture, Bu Ali Sina University, Hamedan, Iran. E-mail: mostafa.rastgo@gmail.com,

[2] **Hossein Bayat:** Associate Professor (Ph. D.), Department of Soil Science, Faculty of Agriculture, Bu Ali Sina University, Hamedan, Iran. Postal Address: Department of Soil Science, Faculty of Agriculture, Bu Ali Sina University, Hamedan, Iran. E-mail: h.bayat@basu.ac.ir Other e-mail: hbayat2001@gmail.com. Office phone: +98-81-34424189, Mobile phone: +98-918-8188378. Fax: +98-81-34424189.

[3] **Muharram Mansoorizadeh:** Assistant professor (Ph. D.), Department of Computer Science, Faculty of Engineering, Bu Ali Sina University, Hamedan, Iran. E-mail: mansoorm@basu.ac.ir

[4] **Andrew S. Gregory**: Sustainable Agriculture Sciences Department, Rothamsted Research, Harpenden, Hertfordshire, AL5 2JQ, UK. E-mail: andy.gregory@rothamsted.ac.uk

*Corresponding author (h.bayat@basu.ac.ir, Other e-mail: hbayat2001@gmail.com)**).**

Estimating the soil water retention curve: comparison of multiple nonlinear regression approach and random forest data mining technique

Abstract

This study evaluates the performance of the random forest (RF) method on the prediction of the soil water retention curve (SWRC) and compares its performance with those of non-linear regression (NLR) and Rosetta-based pedotransfer functions (PTFs), which has not been reported so far. Fifteen RF and NLR-based PTFs were constructed using readily-available soil properties for 223 soil samples from Iran. The general performance of RF and NLR-based PTFs was quantified by the integral root mean square error (*IRMSE*), Akaike's information criterion (*AIC*) and coefficient of determination (R^2). The results showed that the accuracy of the RF-based PTFs was significantly ($P<0.05$) better than the NLR-based PTFs, and ~~also,~~that the reliability of the NLR-based PTFs was significantly ($P<0.01$) better than the RF-based PTFs and all of the Rosetta-based PTFs. The average values of the *IRMSE*, *AIC* and R^2 of the RF method were 0.041 cm^3 cm^{-3}, -16997.7, and 0.987, and 0.053 cm^3 cm^{-3}, -15547.5, and 0.981 for the training and testing steps of all PTFs, respectively, whereas the~~se~~ values for the NLR method were 0.046 cm^3 cm^{-3}, -16616.4, and 0.984, and 0.048 cm^3 cm^{-3}, -16355.6, and 0.983 for the training and testing steps, respectively. The PTF5 of the RF and NLR methods, with ~~the~~ inputs of sand and clay contents, bulk density, and the water content at field capacity and permanent wilting point, had the greatest R^2 values (0.987 and 0.989, respectively), and the lowest *IRMSE* values (0.039 and 0.032 cm^3 cm^{-3}, respectively)~~, respectively,~~ compared to other PTFs for the testing step. Overall, the RF method had less reliability for the prediction of the SWRC compared to the NLR method due to overprediction, uncertainty of determination of forest scale and instability in the testing step. ~~It seems that t~~These findings could provide the scientific basis for further research on the RF method.

Keywords: pedotransfer functions; soil water retention curve; soil texture; soil structure; van Genuchten.

Notation

Sand content (%)	S
Clay content (%)	C
Geometric mean diameter (mm)	d_g
Geometric standard deviation (-)	δ_g
Bulk density (g cm^{-3})	BD
Total porosity (cm^3 cm^{-3})	TP
Water content at field capacity, 33 kPa (cm^3 cm^{-3})	θ_{FC}
Water content at 1500 kPa (cm^3 cm^{-3})	θ_{PWP}
Organic matter content (%)	OM
Saturated hydraulic conductivity (cm day^{-1})	K_s
Saturated water content (cm^3 cm^{-3})	θ_s
Residual water content (cm^3 cm^{-3})	θ_r
Random forest	RF
Nonlinear regression	NLR
Soil water retention curve	SWRC

1 Introduction

Soil hydraulic properties are principle factors that control the movement of water and solutes in the soil. Determination of the soil hydraulic properties is required for many distinct applications linked with irrigation, land use planning, drainage and drought risk assessment (Dobarco et al., 2019). The soil water retention curve (SWRC) is one of the most important soil hydraulic properties. It defines the relationship between soil matric potential and soil water content (Hillel, 1998). The SWRC is a crucial parameter in soil and water management for sustainable and improved agricultural production (Shwetha and Varija, 2015). The SWRC depends principally on texture, structure and bulk density (BD) of soils (Wassar et al., 2016). Many methods have been introduced for the direct measurement of the SWRC in the laboratory (e.g., the hanging water column and pressure plate methods) (Klute, 1986) and in the field (e.g., tensiometric) (Bruce and Luxmoore, 1986). Measurements of the SWRC at several matric potentials can be expensive, difficult and time-consuming, hence it is common to predict it by modelling (Dobarco et al., 2019). Modelling of soil water is an essential tool in evaluating the effects of different managements on crop yield and environmental quality (Verhagen, 1997).

Pedotransfer functions (PTFs) translate easy-to-measure data that we have (e.g., texture class, particle size distribution (PSD) and BD) into difficult-to-measure data that we need (soil hydraulic data) (Bouma, 1989). Estimates of the SWRC by ~~the~~ PTFs are valuable in many studies, such as hydrology, soil mapping and hydrogeology (Børgesen and Schaap, 2005). The point- and parametric-based PTFs are generally developed to predict water content at ~~certain~~ specific matric potential values and the entire SWRC, respectively, by multiple linear (MLR) and nonlinear regression (NLR) methods (Gunarathna et al., 2019b; Merdun et al., 2006; Minasny et al., 1999; Rajkai et al., 2004; Tomasella et al., 2000). Data mining techniques including artificial neural networks (ANNs) (Bayat et al., 2013a; Bayat et al., 2013b; Gunarathna et al., 2019a; Koekkoek and Booltink, 1999; Pachepsky et al., 1996), group method of data handling (GMDH) (Bayat et al., 2011; Neyshaburi et al., 2015; Pachepsky and Rawls, 1999), nonparametric nearest neighbor technique (Botula et al., 2013; Gunarathna et al., 2019a; Haghverdi et al., 2015; Nemes et al., 2006; Nguyen et al., 2017) and support vector machine (SVM) (Khlosi et al., 2016;

Lamorski et al., 2008; Lamorski et al., 2014; Twarakavi et al., 2009), have been applied successfully ~~applied~~ for PTF development.

Random forest (RF), or random decision forests, has become a popular approach as an ensemble learning method for prediction and classification (Verikas et al., 2011). The RF method has been developed by Breiman (2001) as an expansion of the classification and regression trees (CART) technique to provide better performance of prediction results (Wiesmeier et al., 2011). So far, few studies have been carried out on the application of the RF method in soil science. ~~For example,~~ Tóth et al. (2014) applied the RF method to analyze the relationship between soil water content at four matric suctions ~~of~~ (0.1, 33, and 1500 kPa, and 150~~000~~ ~~k~~MPa) and Hungarian soil map information. They found that the importance of soil properties in the prediction of the soil water content varied~~s,~~ according to soil type and matric suction~~s~~. Recently Szabó et al. (2019) have developed PTFs based on RF and geostatistics methods to map soil hydraulic properties, such as water contents at saturation, field capacity and wilting point, for the Balaton catchment area in Hungary. Araya and Ghezzehei (2019) compared the performances of four machine-learning algorithms including -the k-nearest neighbors (kNNs), support vector regression (SVR), RF, and boosted regression tree (BRT) for prediction of ~~the~~ saturated hydraulic conductivity. They found that the BRT model~~s~~ outperformed the other algorithms closely followed by the RF model~~s~~. Gunarathna et al. (2019a) tested three machine~~-~~-learning algorithms including ~~artificial neural networks (~~ANN~~)~~, kNN, and RF to estimate volumetric water content at ~~the~~ matric suctions of 10, 33 and 1500 kPa for soils in Sri Lanka~~n soils~~. They recommended that the PTFs to be developed using the RF algorithm. Ließ et al. (2012) studied uncertainty in the spatial prediction of soil texture by comparison of the RF and regression tree techniques for 56 soil profiles~~. Those authors indicated~~ and found that the ~~RF~~former method provided a better result~~s better than the~~

~~regression tree~~. Also, Wiesmeier et al. (2011) utilized the RF technique to develop digital mapping of the soil organic matter content in 120 soil profiles. They ~~pointed out~~found that the prediction accuracy of the RF modeling was acceptable. A review of literatures therefore revealed that the RF data mining technique has been ~~only~~ applied to develop PTFs to predict specific points~~-based PTFs~~ of the SWRC, such as ~~including~~ field capacity and permanent wilting point, or particular properties such as saturated hydraulic conductivity, but it has not been used ~~for~~to developi~~ng~~ parametric-based PTFs of the van Genuchten model parameters, so far~~The RF data mining technique has not been applied to predict the SWRC, so far~~. Therefore, the objective of the present study was to develop simple parametric-PTFs to predict the SWRC with greater accuracy and reliability using a novel approach with the RF data mining technique. We ~~and~~ compare its performance with those of the multiple ~~non-linear regression (~~NLR~~)~~ approach and with Rosetta software (Schaap et al., 2001) on the prediction of the SWRC through finding the best input variables and PTFs for the SWRC.

2 Materials and methods

2.1 Sample collection and determination

In the present study 223 undisturbed and disturbed soil samples were taken from six provinces of Iran including west Azarbaijan (35°~~°~~ 8□ − 39°~~°~~ 46□ N, 44°~~°~~ 3□ − 47°~~°~~ 23□ E; 60 data), Hamedan (33°~~°~~ 59□ − 35°~~°~~ 48□ N, 47°~~°~~ 34□ − 49°~~°~~ 36□ E; 55 data), Kermanshah (33°~~°~~ 41□ − 35°~~°~~ 17□ N, 45°~~°~~ 24□ − 48°~~°~~ 6□ E; 26 data), Kurdistan (34°~~°~~ 45□ − 36°~~°~~ 31□ N, 45°~~°~~ 31□ − 48°~~°~~ 13□ E; 22 data), Mazandaran (35°~~°~~ 46□ − 36°~~°~~ 58□ N, 50°~~°~~ 21□ − 58°~~°~~ 08□ E; 30 data) and Fars (27°~~°~~ 2□ − 31°~~°~~ 42□ N, 50°~~°~~ 42□ − 55°~~°~~ 38□ E; 30 data). Steel cylinders, measuring 5.1 cm in diameter and 3.5 cm in height, were used to collect the undisturbed samples. Since the sampling was done from different locations of the various provinces, the topsoil and subsoil layers of soil at different locations had different depths and thicknesses. We collected ~~, and~~ samples ~~were taken~~ from the center of the topsoil and subsoil ~~each~~ layers, which represented ~~(”topsoil” and ”subsoil” refer to~~the pedological A and B horizons, respectively~~)~~. ~~Therefore, the samples were taken from different depths, depending on the thickness of the A and B layers.~~ The sampling depths varied from 10 to 35 cm for topsoil (~~A horizon,~~ 208 samples) and from 20 to 45 cm for subsoil (~~B horizon,~~ 15 samples), reflecting the variation in the soil profiles.

Soil PSD was analyzed by the hydrometer method (Gee and Or, 2002), and the geometric mean and standard deviation of particle~~s~~ diameter (d_g and δ_g, respectively) were calculated by equations from Shirazi and Boersma (1984). Organic matter (OM) content was determined by the Walkley and Black (1934) method and BD by the core method (Blake and Hartge, 1986). Total porosity (TP) was calculated from BD and particle density, and the saturated hydraulic conductivity (K_s) was measured with a constant head permeameter (Klute and Dirksen, 1986). The SWRC was ~~conbstructed~~constructed by measuring the volumetric water content at ~~the~~ matric suctions of 0 (saturation status of soil samples), 1, 2 and ~~5 kPa~~5 kPa with a sandbox apparatus, and at 10, 25, 50, 100, 200, 500, 1000 and 1500 kPa with a pressure plate apparatus. Undisturbed samples were used for measurement of the matric suctions from 0 to 100 kPa and disturbed samples were used for matric suctions from 200 to 1500 kPa. Two key points in the SWRC are the water contents at field capacity (30 kPa suction; θ_{FC}) and permanent wilting point (1500 kPa suction; θ_{PWP}).

2.2 Soil-water retention equation

The van Genuchten–Mualem (Eq. (1)) model (Mualem, 1976; van Genuchten, 1980) was utilized to describe the SWRC data.

$$\theta = \theta_r + \left(\theta_s - \theta_r\right) \times \frac{1}{\left[1 + \left(\alpha h\right)^n\right]^{\left(1-\frac{1}{n}\right)}} \qquad (1)$$

where θ_r and θ_s are residual and saturated water contents (cm^3 cm^{-3}), respectively, and h is the soil water suction (kPa). The parameter α is related to the inverse of the air entry pressure (>0, kPa^{-1}) and n (>1, dimensionless parameter) is related to the pore size distribution of the soil (van Genuchten, 1980). In the present study, van Genuchten model parameters θ_r, θ_s, α and n were obtained using the MATLAB software (MathWorks, 2018).

2.3 Data pre-processing

Data pre-processing and regression assumptions, including detection of outliers, normality test of the residuals, multicollinearity and independence of the residuals, were applied for all variables (Berry, 1993). The outliers in the data were identified by the inter-quartile range (IQR) method (Seo, 2006) and were replaced by the lower and upper threshold values (MathWorks, 2018). Before developing PTFs, all variables were evaluated by Kolmogorov-Smirnov normality and multicollinearity tests by the SPSS 24 software (IBM, 2016). The degree of multicollinearity in the PTFs was tested by the variance inflation factor ($VIF=1/1-R^2_j$, where R^2_j is the R^2 value obtained by regressing the j^{th} predictor on the remaining predictors) (Hocking, 2013)~~ (Table 1)~~. ~~The *VIF* values in Table 1 showed low levels of multicollinearity among the independent variables (*VIF*<10) (Khodaverdiloo et al., 2011).~~ Also, t~~T~~o avoid multicollinearity between textural contents, the silt fraction was ~~eliminated~~not used as a predictor. The variables clay content, sand content, d_g, δ_g, OM, K_s, α and n had non-normal distributions, therefore, transformations were applied to normalize them.

2.4 Developing PTFs

The PTF inputs were arranged in four steps (Fig. ~~2~~1). The first step (PTFs 1-5) was based on basic soil properties (i.e., sand content (%), clay content (%), BD (g cm^{-3}), θ_{FC} (cm^3 cm^{-3}) and θ_{PWP} (cm^3 cm^{-3})) according to Rosetta-based PTFs (Schaap et al., 2001) for comparison of SWRC estimates by other methods. ~~To avoid multicollinearity between textural contents, the silt fraction was eliminated.~~ The parameters of the van Genuchten model were predicted in all steps. In the second step (PTFs 6-9), d_g (mm) and δ_g were used as new inputs instead of sand and clay contents in the previous step to evaluate the efficiency of using statistical descriptors of PSD to predict the parameters of the van Genuchten model. To build the third step (PTFs 10-12), TP (cm^3 cm^{-3}) replaced BD from PTFs 3-5 to evaluate the effect of using TP instead of BD on the prediction of the parameters of the van Genuchten model. In other words, the purpose of the second and third steps was to evaluate whether the use of another form of descriptors of ~~the ~~soil structure (TP instead of the BD) and soil texture (d_g and δ_g instead of the sand and clay contents) would improve the accuracy of the estimates or not. In the last step, PTFs 13-15 were developed by including OM (%) and K_s (cm day^{-1}) as new variables to evaluate the efficiency of these instead of the water content at specific matric suctions on the prediction of the van Genuchten model parameters. The input variables of the 15 PTFs are shown in Fig. ~~2~~1.

To compare the results of PTFs 1-5 of the RF and NLR methods with those of the Rosetta models, the parameters of the van Genuchten model (θ_r, θ_s, α and n) were estimated by the PTFs built in the Rosetta software (PTFs 1-5), using the measured values of input variables based on PTFs 1-5 as predictors in the Rosetta program. The estimated coefficients ~~of the~~of the van Genuchten model were used to calculate the estimated water content at matric suctions from 0 to 1500 kPa (estimated SWRCs). Then curve-by-curve comparison of the measured and estimated SWRCs was performed with different evaluation statistics. Since there is no training step in the Rosetta software, the results of the Rosetta model was only compared with the results of the testing step. To evaluate the effect of using different descriptors of PSD on the prediction of the SWRC, PTFs 6, 7, 8 and 9 from the second step were compared with PTFs 2, 3, 4 and 5 from the first step, respectively (Fig. ~~2~~1). In the same way, to evaluate effect of using different descriptors of soil structure on the prediction of the SWRC, PTFs 10, 11 and 12 from the third step were compared with PTFs 3, 4 and ~~-~~5 from the first step, respectively. Also, the PTFs 13-15 were compared with the PTFs 4 and 5 to find out the efficiency of OM and K_s variables as predictors (Fig. ~~2~~1).

Fig ~~2~~1.

In the present study, the k-fold cross validation approach (Efron and Tibshirani, 1994) was utilized to obtain training and testing datasets for each PTF. The number of folds (i. e., k) ~~h~~was ~~been~~ obtained by trial and error. To do so, some PTFs, ~~which were~~ selected randomly, ~~have been~~were developed with 10, 15 and 20-fold cross-validation. Then, the k value which ~~was~~ resulted in the best performance of the PTFs, was selected to develop all PTFs in this study. The results showed that 20-fold cross validation performed better than the other folds~~,~~ in most of the PTFs (Table 1). Therefore, 20-fold cross validation was selected to develop PTFs in this study. Based on this approach, the 223 samples were randomly divided into 20 subsets and 20 models were developed by each predicting technique for each PTF. In each model, training and testing datasets were based on a ratio of 19:1. Finally, the average of the results of 20 models was calculated for each PTF. Therefore, all data were used for the training and testing steps of the PTFs.

Table 1-

2.5 Description of modeling techniques

2.5.1 Multiple nonlinear regression

A ~~nonlinear regression~~NLR model based on a second-order polynomial for the prediction of the response variable *y* from a number of ~~*n*~~ *p* predictors can be written as (Rawls and Brakensiek, 1985):

$$y = a + \sum_{i=1}^{p} \left(b_i x_i + c_i x_i^2 \right) \qquad (2)$$

where a is the intercept, and two regression coefficients b_i and c_i are determined for every input variable x_i.

2.5.2 *Random forest: an ensemble of regression trees*

RF has become a popular tool for regression and classification problems. The RF is an ensemble method based on the regression tree methodology (i.e., ~~classification and regression trees (~~CART~~)~~) that was introduced for better performance (Breiman, 2001). The model building process in the RF is the same as that in the CART method but without pruning (Breiman, 1984). Also, whereas a regression tree only grows by a single tree~~, but~~ the RF grows by forest of trees. In other words, unlike a regression tree, in the RF for each tree only a subset of the input variables is applied. The number of inputs in each tree and also the number of trees in the forest can be distinct and it depends on the dataset. Least-squares boosting (LSBoost) fits regression ensembles. At every step, the ensemble fits a new learner to the difference between the observed response and the aggregated prediction of all learners grown previously. The ensemble fits to minimize the mean-squared error (MathWorks, 2018). The number of trees used here was 16 which was established by trial and error. An architecture of the RF algorithm is shown in Fig. ~~3~~ 2 where input matrix X consists of N samples and M input variables (sample set S = [(x_i, y_i), i = 1, 2, …, N], (X, Y~~)∈~~) ∈R^M×R). The bootstrap method is utilized to construct n tree sample sets from the sample set S. At each bootstrap sample, about one-third of the dataset S was utilized as out of the bootstrap data or out-of-bag (*OOB*) data; whereas the rest is called in-bag data (Ibrahim and Khatib, 2017) (Fig. ~~3~~2). Modeling of the regression tree is done for each sample set. In the RF algorithm, all individual trees give a predictive result. The final prediction value is calculated based on an average result of all individual trees (Wiesmeier et al., 2011). The prediction error is defined as follows (Liaw and Wiener, 2002):

$$MSE_{OOB} = \frac{\sum_{i=1}^{n_{tree}} \left(y_i - \hat{y}_i^{OOB} \right)^2}{n_{tree}} \tag{3}$$

where MSE_{OOB} is the mean square error of the *OOB* data prediction, n_{tree} is the number of trees, and y_i and $\hat{y}_i^{OOB}$ are the actual value of the *OOB* data and the average of all *OOB* predictions, respectively. Among all the ensemble methods, the RF method has high capability in solving classification and regression problems, because the RF method combines several simple regression trees to better optimize prediction (Zaklouta and Stanciulescu, 2012). The RF method increases differences for each single tree through random selection of the training samples and different variables at each splitting node. In the present study, the NLR and RF algorithms were implemented by fitnlm and fitensemble functions in the MATLAB software, respectively. (MathWorks, 2018).

Fig. ~~3~~2.

2.6 Evaluation criteria

The estimated water content was computed by estimated parameters of the van Genuchten model for each PTF at matric suctions from 0 to 1500 kPa. For curve-by-curve comparison of the measured and predicted SWRCs, different evaluation statistics were used. Various statistical criteria including integral root mean square error (*IRMSE*), integral mean error (*IME*) (Tietje and Tapkenhinrichs, 1993), Akaike's information criterion (*AIC*) (Akaike, 1974) and coefficient of determination (R^2) (Wösten et al., 2001), were utilized to assess the predictive ability of the RF and NLR algorithms, which are defined as:

$$IRMSE\left(cm^{3}cm^{-3}\right)=\left[\frac{1}{b-a}\int_{a}^{b}(\hat{y}_{i}-y_{i})^{2}\,d\,\log|h|\right]^{\frac{1}{2}} \qquad (4)$$

$$IME\left(cm^{3}cm^{-3}\right)=\frac{1}{b-a}\int_{a}^{b}(\hat{y}_{i}-y_{i})\,d\,\log|h| \qquad (5)$$

$$AIC=N\times\ln\left[\sum_{i=1}^{N}\frac{(y_{i}-\hat{y}_{i})^{2}}{N}\right]+2P \qquad (6)$$

$$R^{2}=1-\frac{\sum_{i=1}^{N}(y_{i}-\hat{y}_{i})^{2}}{\sum_{i=1}^{N}(y_{i}-\bar{y}_{i})^{2}} \qquad (7)$$

where *h* is the matric suction (kPa)~~,~~, y_i, $\hat{y}_i$ and $\bar{y}_i$ are the measured, predicted and average of the measured values of the water content, respectively, *a* and *b* values define the matric suction range over which the experimental curve is measured, i.e., 0 and 1500 kPa, respectively, and *P* and *N* are the number of parameters and the number of points that were considered in the SWRC, respectively. In calculating the *AIC*, *N* is the total number of points that were considered in the SWRC of all soil samples (i. e., *N*= number of soil samples × number of paired points of the suction-water content for each soil sample), and *i* is paired points of the suctions-water content of the SWRC of each soil sample.

To evaluate the performance of each method in different PTFs, the effect of method~~s~~ as the first factor at two levels~~,~~ in the training step (*i.e.*, NLR and RF methods) ~~in the training step~~ and at three levels~~,~~ in the testing step (*i.e.*, NLR, RF and Rosetta methods)~~ in the testing step~~, and the different PTFs as the second factor at 15 levels (PTF1 to PTF15), were investigated using a two-way analysis of variance (ANOVA) with a randomized complete block design~~ as a factorial test~~, based on the *IRMSE* of prediction of the SWRC. ~~On the other hand, t~~The *IRMSE* criterion calculates the total error, including bias and random errors, and is a more appropriate criterion for evaluating the accuracy and reliability of the RF and NLR methods compared to other criteria (Chai and Draxler, 2014). Therefore, t~~T~~o compare the predicting accuracy and reliability of the RF and NLR methods, the average values of the *IRMSE* was compared with Duncan's test by MathWorks (2018) software.

3 Results and discussion

3.1 Descriptive statistics of the soil properties

Table ~~1~~ 2 summarizes some basic descriptive statistics for soil variables of the entire dataset used for the development of the PTFs. It can be seen that the average and maximum of clay content were 21.4 and 48%, respectively.~~It can be seen that the average clay content was 21.4 %, and exceeded 50%.~~ The OM ranged from 0.17 to 4.41% with a mean of 1.84%, which ~~i~~was low due to the arid and semi-arid climates of Iran. The variation ~~of the~~in soil texture is shown graphically in the United States Department of Agriculture (USDA) textural triangle (Fig. ~~4~~3). Considering the distribution and range of the variables (Fig. ~~4~~ 3 and Table ~~1~~2), the dataset can be considered as representative of soils in arid and semi-arid regions of Iran.

Table ~~1~~2

Fig. ~~4~~3.

3.2 Correlation of input and output variables

The simple correlation coefficients between all variables are depicted by matrix plot in Fig. ~~1~~4. Correlation analysis was done between normalized input and output variables. The correlation test was not performed for the θ_r variable, because its value was zero in 138 out of 223 soil samples,~~.~~

~~Also the zero value~~ as has~~ve~~ been reported in ~~some~~ other studies (Campbell and Horton Jr, 2002; Rawls et al., 1991; Tomasella et al., 2000) for θ_r variable. Clay and sand contents, θ_{FC}, θ_{PWP}, d_g and OM had the greatest significant correlations with the parameters of the van Genuchten model (Fig. ~~1~~4), which ~~are~~ was consistent with other studies (Dexter et al., 2008; Nemes et al., 2006). For example, the correlation coefficient~~s~~ between clay content and θ_s ($r = 0.323$) is close to that between ~~the~~ OM and θ_s ($r = 0.268$). Also, the results showed that there were significant correlations between θ_{PWP} and input variables of clay content (+), sand content (~~-~~—), BD (~~-~~—), OM (+) and K_s (~~-~~—), and also between θ_{PWP} and θ_s (+) and n (~~-~~—) parameters of the van Genuchten model (Fig. 4). Botula et al. (2012) also found the same observation for the correlation of θ_{PWP} with sand and clay contents and BD of tropical Lower Congo soils. Nevertheless, with regard to these correlation coefficients, clay and sand contents, θ_{FC}, d_g and OM can be used for developing PTFs to estimate the SWRC. On the contrary, there was no correlation between K_s and the van Genuchten model parameters. There are many cases, where two variables might not show a strong simple correlation, but may show a strong association in the regression, along with other predictors. In other words, the simple correlation coefficient is a way to show the relationship between ~~two~~ independent and dependent variables, but it cannot show a model for the relationship between these two variables, when other independent variables have been used in a multiple regression (Simmons et al., 2011). The result~~s~~ of multiple regression analysis with backward selection method showed that the K_s variable remained in the PTF14 and PTF15 for all the van Genuchten model parameters. Some of the regression equations with backward selection method are shown in the following as examples:

$$\theta_r=-0.69+0.22\times Clay+0.278\times Sand+0.20\times K_s,\ R=0.31^{**} \qquad (8)$$

$$\alpha=-3.72+0.23\times Clay+0.17\times BD+0.282\times K_s,\ R=0.33^{**} \qquad (9)$$

$$n=-1.76+0.24\times \text{Sand}+0.164\times K_s,\ R=0.30^{**} \qquad (10)$$

On the other hand, the non-linear correlations between variables are very important in this study. B~~ecause, b~~oth the multiple ~~nonlinear regression~~NLR approach and ~~random forest~~RF data mining technique~~, which were used,~~ are non-linear prediction methods. Fig. 4 only shows simple linear correlation between variables, but there may be non-linear correlations between variables, which may affect the estimation of the dependent variables. For example, the results of non-linear correlations showed that K_s had strong correlations with θ_s and α of the van Genuchten model parameters by logarithmic ($\theta_s=0.652-0.027\times \ln K_s$, $R=0.62^{**}$) and power ($\alpha=0.007\times K_s^{0.283}$, $R=0.57^{**}$) equations, respectively, ~~and these non-linear correlations~~which were ~~increased mostly in comparison with~~greater than their simple correlations~~, indicating nonlinear relationships of the K_s with θ_s and α. Therefore, regression method can discover and apply the law that exists between these two variables.~~

Fig. 4~~1~~.

3.3 Development of the PTFs using the RF and NLR methods

Results of the multicollinearity analysis (*VIF*) are shown in Table ~~2~~3. The *VIF* values ~~in Table 2~~ showed low levels of multicollinearity among the independent variables (*VIF*<10) (Khodaverdiloo et al., 2011).

Table ~~2~~3-

3.3.1 Comparing the accuracy and reliability of the RF and NLR methods

Table ~~3~~4 shows the results of ~~analysis of variance~~the ANOVA of the *IRMSE* of prediction of the SWRC by different methods and PTFs. The ~~analysis of variance showed that the~~ effect of ~~type of~~ methods and PTFs, and their interaction, on the *IRMSE* was significant at *P*-<-0.01, 0.01 and 0.05, respectively, in the training step, and ~~also~~at *P*-<-0.01, 0.01 and 0.01, respectively, in the testing step. Therefore, we focus on the results and discussion of the~~mean~~ comparison ~~was performed and results and discussion were written according to~~of the method × PTF interaction effects.

Table ~~3~~4-

Results of the prediction of the SWRC through the van Genuchten model using the NLR and RF-based PTFs are depicted in Figs. 5 and ~~6 for~~6 for the training and testing steps, respectively. The accuracy and reliability are used to ~~express the~~express the performance of the PTFs in the training and testing steps, respectively.

The results of the first to fourth steps of the training dataset (Fig. 5) showed that the RF method had better performance compared to the NLR method for the prediction of the SWRC in all PTFs in terms of the *IRMSE* and R^2 criteria and the differences were significant (*P*-<-0.05) for PTFs 2, 3, 6, 7, 10, 13, 14 and 15 in terms of the *IRMSE* criterion. Also, the accuracy of the RF method was better than that of the NLR method in 80% of the PTFs (with the exception of the PTFs 5, 9 and 12) in terms of the *AIC* criterion. In the training step, the values of the *IRMSE* of the first to fourth steps for the NLR model varied from 0.030 to 0.063 cm^{3} cm^{-3} and these were larger than those in the RF model, which ranged from 0.028 to 0.061 cm^{3} cm^{-3}, respectively. Also, the values of the R^2 of the first to fourth steps for the RF model varied from 0.981 to 0.992, and this was larger than those in the NLR model, which ranged from 0.979 to 0.991 (Fig. 5).

The results of the first to fourth steps of the testing dataset (Fig. 6) showed that the NLR method had a better performance compared to the RF method on the prediction of the SWRC for PTFs 5, 8, 9 and 15 only in terms of the *IRMSE* criterion (significant at $P<0.05$). In the other PTFs there were no significant differences between the *IRMSE* of the two methods and the R^2 and *AIC* criteria were comparable. In the testing step, the values of the *IRMSE* and *AIC* of the first to fourth steps for the RF models varied from 0.038 to 0.065 cm^3 cm^{-3} and from -13476.2 to -17646.8, respectively, and these were comparable to those of the NLR models (with the exception of ~~the~~ PTF1), which ranged from 0.032 to 0.064 cm^3 cm^{-3} and from -14096.1 to -19234.1, respectively (Fig. 6). Also, the values of the R^2 of the first to fourth steps for the NLR models varied from 0.979 to 0.989, and this was comparable to those of the RF models for all PTFs, which ranged from 0.977 to 0.987 (Fig. 6).

In each of the PTFs 1 to 5, the NLR and RF methods performed better ($P<0.05$) than the Rosetta PTFs. Fig. 6(A) shows that the Rosetta-based PTFs ~~have~~ had greater values of the *IME* criterion compared to the NLR and RF-based PTFs. The reason can be attributed to the various methods of optimizing parameters. The Rosetta method has only one ~~artificial neural network (~~ANN~~)~~ type with particular structure. In other words, the number of hidden layers (one) and neurons (six) and also the activation function (tangent hyperbolic) are constant for prediction of the SWRC in the Rosetta software. Therefore, the Rosetta method is not a dynamic approach for optimization, whereas the parameters of the RF method, such as number of splits and trees, and learning rate continuously and dynamically, change to achieve the best result of the objective function. The Rosetta method was developed from a ~~quite~~ large dataset, while the soils used in the present study were collected from a completely different climate area that was not represented in the Rosetta's database. Also, presented RF and NLR models were trained using this particular dataset while Rosetta had been trained using a different dataset. In other words, the results of the PTFs in the testing step ~~are~~ were based on a soil dataset used for training. This could be a reason for Rosetta's poor performance compared with the RF and NLR methods. As a result, it seems that the universal portability of the Rosetta method can be limited. The testing results are in agreement with Touil et al. (2016) who found that the parametric-based PTFs of nonlinear models~~,~~ gave a better prediction than the Rosetta PTFs. The Figs. 5(A) and 6(A) showed that all of the *IME* values were negative for all PTFs at the training and testing steps. There are regular errors (bias) in the prediction of the SWRC that can be corrected by finding a correction coefficient, which would improve the accuracy and reliability of the estimations (Bayat et al., 2015).

Fig. 5.

Fig. 6.

The RF method in the training section gave better predictions of the SWRC compared to the NLR method ~~for the prediction of the SWRC~~ (Fig. 5). The RF method produces low~~-~~ bias and variation ~~results~~ in the data by majority voting compared to a single regression tree (Cheng et al., 2019; Matin and Chelgani, 2016). In this connection, the results of the standard deviations (SD) of evaluation criteria in each PTF for the training step (Fig. 5) showed that the RF method had a lower ~~SD~~ variation than the NLR method. Accordingly, the values of SD for the *IRMSE* and R^2 criteria were 0.024 and 0.022, respectively, for the NLR model~~,~~ and these were larger than those in the RF model, which were 0.020 and 0.017, respectively, for the training step. On the other hand, the RF method can be applied to high dimensional datasets in regressions (Janitza et al., 2016; Zhao et al., 2016).

As depicted in Fig. 6, unlike in the training section, the NLR method gave better predictions in the testing section~~,~~ compared to the RF method for the prediction of the SWRC. In other words, the reliability of the NLR method was better than that of the RF method in all the PTFs. The ~~nonlinear regression~~NLR equations can be more useful than the MLR method for the prediction of the SWRC due to their high flexibility (Williams et al., 1992). In other words, the NLR models have capacity to capture nonlinear relationships in the dataset. Tomasella et al. (2000) successfully developed parametric~~-~~ PTFs for soils of the humid tropics using polynomials of nth order. Medrado and Lima (2014) successfully developed NLR-based PTFs to predict the four parameters of the van Genuchten model for Brazilian soils. Also, Touil et al. (2016) developed parametric-PTFs to predict the SWRC using the NLR method from more readily-available properties such as soil texture, OM content, and BD for 242 soil samples of Algeria. They reported that the parametric-PTFs had better performance ~~compared to the~~than Rosetta-based PTFs.

In the present study, in contrast to the NLR method~~,~~ which had less differences between the error values of the training and testing steps~~, the~~, the error values of the RF method in the testing dataset ~~was~~ were much greater than those in the training dataset. These results can be due to overprediction phenomenon in the RF method. Gupta et al. (2017) expressed that one of the disadvantages of the RF method is the overprediction. In other words, the RF method is a 'greedy' method that easily leads to overprediction and instability in the testing step and solving this problem can be of great significance for improving the reliability of the RF method (Liu, 2014). Also, Ma et al. (2005) reported ~~the~~ instability in results of the RF method. The forest size developed by the RF has not been clearly defined (Liu, 2014). Therefore, oversized scale can decrease the reliability and efficiency of the SWRC prediction. Hong et al. (2016) evaluated landslide susceptibility maps produced using the RF method and compared these maps with those from statistical-based methods, such as logistic regression, and their study revealed that the performance of the statistical-based methods was better than that of the RF method. ~~Also, a~~A similar result was reported by Esposito et al. (2014). Generally, RFs are best suited for problems with many input variables and a reasonable sample size. According to ~~the~~ our results (Figs. 5 and 6), performance of the PTFs was improved by increasing the number of input variables.

3.3.2 Evaluation of the effect of the basic soil properties on prediction performance of the SWRC

A significant improvement was achieved in the accuracy of PTF5 (with the inputs of Sand content+Clay content+BD+θ_{FC}+θ_{PWP}) compared to other PTFs (with the exception of PTFs 4, 8, 9, 11 and 12) by both NLR and RF methods in terms of the *IRMSE* criterion (Fig. 5). Among the PTFs of each method (RF or NLR), PTF5 had the greatest R^2 (0.992 and 0.991, respectively) and the smallest *IRMSE* (0.028 and 0.03, respectively) and *AIC* (-19432 and -19571.1, respectively) ~~values,~~ in the training step of the prediction of the SWRC. In connection with the importance of input variables, an improvement was achieved in the reliability of the prediction of the SWRC by PTFs 9 (with the inputs of d_g+δ_g+BD+θ_{FC}+θ_{PWP}) and 12 (with the inputs of Sand content+Clay content+TP+θ_{FC}+ θ_{PWP}) from the second and third steps, using the NLR (*IRMSE*=0.032 cm^3 cm^{-3}, *AIC*=-19234.1 and R^2=0.989) and RF (*IRMSE*=0.038 cm^3 cm^{-3}, *AIC*=-17646.8 and R^2=0.987) methods, respectively, in comparison with the other PTFs of each method (Fig 6). However, the differences of ~~the~~ PTFs 9 and 12 were not significant (*$P<0.05$*) with PTFs 4, 5, 8, 11 and 12 in the NLR method and ~~also~~ with PTFs 4, 5, 8, 9 and 11 in the RF method, respectively, in terms of the *IRMSE* criterion.

3.3.2.1 Effect of using different input variables of PSD and soil structure as predictors on the SWRC prediction

~~Input variables such as textural contents (clay and sand contents) and statistics (d_g and δ_g) as different descriptors of the PSD, and also the TP and BD as different descriptors of the soil structure, were used for prediction of the SWRC. Thus, t~~To evaluate the effect of using different descriptors of the PSD on the prediction of the SWRC, PTFs 2, 3, 4 and 5 (clay and sand contents) from the first step were compared with PTFs 6, 7, 8 and 9 (d_g and δ_g) from the second step, respectively. In the same way, to evaluate the effect of using different descriptors of ~~the~~ soil structure on the prediction of the SWRC, PTFs 3, 4 and 5 (BD) were compared with PTFs 10, 11 and 12 (TP) from the third step, respectively. The accuracy and reliability of the prediction of the SWRC by both NLR and RF methods were not significantly different ($P<0.05$) (Figs. 5B and 6B). For descriptors of soil structure, the accuracy and reliability of the prediction of the SWRC by both NLR and RF methods decreased in terms of the *IRMSE* criterion for PTFs 10 to 12 from the third step compared to PTFs 3 to 5 (with the exception of PTFs 11 and 12 in the testing step for the RF method), respectively, when TP was used instead of BD in the list of input variables (Figs. 5B and 6B). However, the differences were not significant ($P<0.05$).

The lack of significant differences between textural contents (clay and sand contents) and statistics (d_g and δ_g), and also between TP and BD on the SWRC prediction can be due to correlation of these parameters with the parameters of the van Genuchten model (Fig. ~~1~~4). The SWRC ~~can be~~is strongly influenced by the soil structure or pore-size distribution and soil texture at small and great matric suctions, respectively (Pachepsky et al., 2006). Therefore, input variables of the textural contents or statistics can influence the residual saturation region of the SWRC.~~,~~ However, soil water content at the dry end (high matric suctions) of the SWRC is primarily determined by textural contents (Hillel, 1998)~~which had significant correlations with θ_r parameter (with the exception of the clay content) (Fig. 1)~~. Also, TP and BD are indicators of soil structure and had significant correlations with θ_s (Fig. ~~1~~4). Indeed, TP was calculated by BD and particle density (Rab et al., 2011).~~On the other hand, t~~The d_g and δ_g predictors were derived from soil textural contents (Shirazi and Boersma, 1984).~~ In other words, textural contents data can be converted to d_g and δ_g by equations of Shirazi and Boersma (1984). Also, TP was calculated by BD and particle density (Rab et al., 2011).~~ Therefore, these could be reasons for similar effects of textural contents and statistics and also TP and BD predictors on the prediction of the SWRC.

Many researchers used textural contents (Adhikary et al., 2008; Chakraborty et al., 2011; Minasny et al., 1999; Tomasella and Hodnett, 1998), d_g and δ_g (Rab et al., 2011; Scheinost et al., 1997; Ungaro et al., 2005), BD (Bayat et al., 2011; Pachepsky et al., 1998) and TP (Bayat et al., 2011; Pachepsky et al., 1998; Schaap et al., 1998) as effective predictors to derive point- and parametric-PTFs. Nemes et al. (2003), Schaap et al. (2001) and Schaap et al. (1998) reported that the variables of PTF5 have better capability on predicting the parameters of the van Genuchten (1980) model with an average *RMSE* of 0.026, 0.044 and 0.058 cm^3cm^{-3}, respectively.

According to the results of the accuracy (Fig. 5) and reliability (Fig. 6) of PTFs 5, 9 and 12, it seems that certain points of the SWRC (e.g., θ_{FC}) can help to improve the prediction of the SWRC and this is in agreement with Schaap et al. (2001). These results indicate that the presence of at least one moisture point~~s~~ (e.g., θ_{FC}) can improve the prediction of the SWRC. In ~~other words, according to the results of the accuracy (Fig. 5) and reliability (Fig. 6) of the NLR and RF methods for different PTFs, at least one moisture point is necessary to predict the SWRC. For example, in~~ the first step, PTF5 with two moisture points (θ_{FC}+θ_{PWP}) and PTF4 with one moisture point (θ_{FC}) improved the prediction of the SWRC by 55, 48, 42% and 51, 44, 38% in terms of the *IRMSE* criterion compared to the PTFs 1, 2 and 3, respectively, in the RF method in the training step. In the testing section of the second step, PTF9 with two moisture points (θ_{FC}+θ_{PWP}) and PTF8 with one moisture point (θ_{FC}) decreased the *IRMSE* by 49, 44% and 44, 39% compared to PTFs 6 and 7, respectively, in the NLR method. The points above are also true for the RF-based PTF12 in the third step of the testing section. Many researchers successfully applied θ_{FC} and θ_{PWP} as effective predictors to derive point- and parametric-PTFs (Børgesen and Schaap, 2005; Nemes et al., 2003; Schaap et al., 2001; Touil et al., 2016; Twarakavi et al., 2009).

3.3.2.2 Effect of using OM and K_s as predictors on the SWRC prediction

To evaluate the effect of using OM and/or K_s and points of the SWRC on the prediction of the SWRC, the performances of PTFs 13, 14 and 15 were compared with those of PTFs 4 and 5. The accuracy and reliability of the prediction of the SWRC by both NLR and RF methods, significantly (*P*<0.05) decreased in terms of the *IRMSE*, for the PTFs 13, 14 and 15 from the fourth step, when OM and/or K_s were used with textural contents and BD as inputs instead of θ_{FC} or both θ_{FC} and θ_{PWP} in the list of input variables, compared to PTFs 4 and 5 at the first step (Figs. 5B and 6B). Therefore OM and K_s were not as effective predictors as θ_{FC} and θ_{PWP} in the prediction of the SWRC, because θ_{FC} and θ_{PWP} are two points of the SWRC and enter direct information of the SWRC into the PTFs, whereas OM and K_s enter indirect information, and therefore had less effect in the improvement of the estimation of the SWRC. These results agreed well with results obtained by Børgesen and Schaap (2005). They reported that PTFs with the inputs of θ_{FC} and θ_{PWP} had smaller *RMSE* values than a PTF with the input of OM (0.038 versus 0.042) in the prediction of the SWRC. On the other hand, the results showed that by adding OM and/or K_s as predictors in the PTFs 13, 14 and 15, the accuracy (Fig. 5B) and reliability (Fig. 6B) of the prediction of the SWRC improved by 16, 13, 17 and 7.1, 6.3, 6.9%, respectively, compared to the PTF3 in terms of the *IRMSE* criterion in the RF method.

The SWRC depends mainly on the soil texture and structure (Hillel, 1998), with OM affecting the SWRC through development of soil structure (Nemes et al., 2005), important at low suctions. However, the OM retains water itself. Similarly, K_s can be a descriptive index of soil texture and porosity (Hillel, 1998). The correlation results showed ~~(Fig. 14)~~ that K_s can be strongly influenced by clay content and textural statistics (d_g and δ_g) ~~soil texture and TP~~(Fig. 4). Bayat et al. (2013b) applied OM and K_s to estimate water content at the measured matric suctions. They found that the OM and K_s can be most appropriately used in point-based PTFs to estimate water content at the matric suctions of 25 and 50 kPa. Also, the result of the present study agreed well ~~to the~~with results obtained by Hollis et al. (1977) and Rawls et al. (1983). In this study, the OM and K_s in the PTFs 13, 14 and 15 were not effective predictors compared to θ_{FC} and θ_{PWP} in the PTFs 4 and 5, otherwise they had better results than PTF3.

4 Conclusion

Machine-learning tools have been widely applied for the prediction of the SWRC. The present study evaluated the capability and performance of the RF method as a novel machine learning tool and compared its performance with that of the ~~nonlinear regression (~~NLR~~)~~ method on the prediction of the SWRC, using different combinations of easily-available soil properties. It was found that the RF method had a better performance (P<0.05) than the NLR method in the training step of the prediction of the SWRC in term of the *IRMSE*, *AIC* and R^2 criteria. However, in the testing step, NLR had a better performance than RF. The poor performance of the RF compared to the NLR method could be due to overprediction in the former, resulting in instability in the testing step. The RF method can be sensitive to sparse areas on the prediction space. In other words, the performance and~~,~~ sensitivity of predictions, and the computational intensity of the RF method depends on the distribution and number of observations and input variables. Therefore, ~~this~~ the method should be tested further with different datasets to evaluate its performance through soil and water investigations. An improvement was achieved in the accuracy of the prediction of the SWRC in the training step of the PTF5 (with the inputs of Sand content+Clay content+BD+θ_{FC} +θ_{PWP}) by both NLR and RF methods and also an improvement was achieved in the reliability of the PTF9 (with the inputs of d_g+δ_g+BD+θ_{FC}+θ_{PWP}) and PTF12 (with the inputs of Sand content +Clay content+TP+ θ_{FC}+θ_{PWP}) by the NLR and RF methods compared to other PTFs, respectively. Considering that the PTFs 5, 9, and 12 had no significant difference from PTF4 (with the inputs of Sand content+Clay content+BD+θ_{FC}) and PTF8 (with the inputs of d_g+δ_g+BD+θ_{FC}+θ_{PWP}), these latter PTFs, with less and more-easily measured input variables, are suggested to be the best PTFs for the prediction of the SWRC. Also, PTFs without predictors of θ_{FC} and θ_{PWP}, such as the PTF3 (with the inputs of Sand content+Clay content+BD) and PTF7 (with the inputs of d_g+ δ_g+BD), can be effective models for the prediction of the SWRC.

Acknowledgements

This work was funded by Bu Ali Sina University, Hamedan, Iran. The authors are deeply grateful to anonymous reviewers and the editor for their helpful comments on the manuscript.

References

Adhikary, P.P., Chakraborty, D., Kalra, N., Sachdev, C., Patra, A., Kumar, S., Tomar, R., Chandna, P., Raghav, D., Agrawal, K., 2008. Pedotransfer functions for predicting the hydraulic properties of Indian soils. Soil Res. 46, 476-484.

Akaike, H., 1974. A new look at the statistical model identification. IEEE transactions on automatic control 19, 716-723.

Araya, S.N., Ghezzehei, T.A., 2019. Using Machine Learning for Prediction of Saturated Hydraulic Conductivity and Its Sensitivity to Soil Structural Perturbations. Water Resour. Res. 55, 5715-5737.

Bayat, H., Ersahin, S., Hepper, E.N., 2013a. Improving estimation of specific surface area by artificial neural network ensembles using fractal and particle size distribution curve parameters as predictors. Environ. Model Assess. 18, 605-614.

Bayat, H., Neyshabouri, M., Mohammadi, K., Nariman-Zadeh, N., 2011. Estimating water retention with pedotransfer functions using multi-objective group method of data handling and ANNs. Pedosphere 21, 107-114.

Bayat, H., Neyshaburi, M.R., Mohammadi, K., Nariman-Zadeh, N., Irannejad, M., 2013b. Improving water content estimations using penetration resistance and principal component analysis. Soil Tillage Res. 129, 83-92.

Bayat, H., Sedaghat, A., Sinegani, A.A.S., Gregory, A.S., 2015. Investigating the relationship between unsaturated hydraulic conductivity curve and confined compression curve. J. Hydrol. 522, 353-368.

Berry, W.D., 1993. Understanding regression assumptions. Sage Publications, London.

Blake, G., Hartge, K., 1986. Bulk density, Methods of Soil Analysis: Part 1. Physical and Mineralogical Methods, Madison, Wisconsin, USA: Soil Sci. Soc. Am. J.

Børgesen, C.D., Schaap, M.G., 2005. Point and parameter pedotransfer functions for water retention predictions for Danish soils. Geoderma 127, 154-167.

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

Botula, Y.-D., Cornelis, W.M., Baert, G., Mafuka, P., Van Ranst, E., 2013. Particle size distribution models for soils of the humid tropics. Journal of Soils and Sediments 13, 686-698.

Bouma, J., 1989. Using soil survey data for quantitative land evaluation, Advances in soil science. Springer, pp. 177-213.

Breiman, L., 1984. Classification and regression trees. Routledge, New York.

Breiman, L., 2001. Random forests. Machine learning 45, 5-32.

Bruce, R.R., Luxmoore, R.J., 1986. Water Retention: Field Methods, In: Klute, A. (Ed.), Methods of Soil Analysis: Part 1—Physical and Mineralogical Methods. Soil Science Society of America, American Society of Agronomy, Madison, WI, pp. 663-686.

Campbell, G.S., Horton Jr, R., 2002. Methods of Soil Analysis: Part 4, Physical Methods. Soil Sci. Soc. Am.

Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)?– Arguments against avoiding RMSE in the literature. Geosci. Model Dev. 7, 1247-1250.

Chakraborty, D., Mazumdar, S., Garg, R., Banerjee, S., Santra, P., Singh, R., Tomar, R., 2011. Pedotransfer functions for predicting points on the moisture retention curve of Indian soils. Indian J. Agr. Sci. 81, 1030.

Cheng, L., Chen, X., De Vos, J., Lai, X., Witlox, F., 2019. Applying a random forest method approach to model travel mode choice behavior. Travel behaviour and society 14, 1-10.

Dexter, A., Czyż, E., Richard, G., Reszkowska, A., 2008. A user-friendly water retention function that takes account of the textural and structural pore spaces in soil. Geoderma 143, 243-253.

Dobarco, M.R., Cousin, I., Le Bas, C., Martin, M.P., 2019. Pedotransfer functions for predicting available water capacity in French soils, their applicability domain and associated uncertainty. Geoderma 336, 81-95.

Efron, B., Tibshirani, R.J., 1994. An introduction to the bootstrap. CRC press.

Esposito, C., Barra, A., Evans, S.G., Scarascia Mugnozza, G., Delaney, K., 2014. Landslide susceptibility analysis by the comparison and integration of random forest and logistic regression methods; application to the disaster of Nova Friburgo-Rio de Janeiro, Brasil (January 2011), EGU General Assembly Conference Abstracts.

Gee, G.W., Or, D., 2002. 2.4 Particle-Size Analysis, In: Dane, J.H., Topp, C.G. (Eds.), Methods of Soil Analysis: Part 4 Physical Methods. Soil Science Society of America, Madison, WI, pp. 255-293.

Gunarathna, M., Sakai, K., Nakandakari, T., Momii, K., Kumari, M., 2019a. Machine Learning Approaches to Develop Pedotransfer Functions for Tropical Sri Lankan Soils. Water 11, 1940.

Gunarathna, M., Sakai, K., Nakandakari, T., Momii, K., Kumari, M., Amarasekara, M., 2019b. Pedotransfer functions to estimate hydraulic properties of tropical Sri Lankan soils. Soil Till. Res. 190, 109-119.

Gupta, B., Rawat, A., Jain, A., Arora, A., Dhami, N., 2017. Analysis of various decision tree algorithms for classification in data mining. Int. J. Comput. Appl. 163, 15-19.

Haghverdi, A., Leib, B.G., Cornelis, W.M., 2015. A simple nearest-neighbor technique to predict the soil water retention curve. Transactions of the ASABE 58, 697-705.

Hillel, D., 1998. Environmental soil physics: Fundamentals, applications, and environmental considerations. Academic press.

Hocking, R.R., 2013. Methods and applications of linear models: regression and the analysis of variance. John Wiley & Sons.

Hollis, J., Jones, R., Palmer, R., 1977. The effects of organic matter and particle size on the water-retention properties of some soils in the West Midlands of England. Geoderma 17, 225-238.

Hong, H., Pourghasemi, H.R., Pourtaghi, Z.S., 2016. Landslide susceptibility assessment in Lianhua County (China): a comparison between a random forest data mining technique and bivariate and multivariate statistical models. Geomorphology 259, 105-118.

IBM, C., 2016. IBM SPSS Statistics for Windows, Version 24.0. Armonk, NY: IBM Corp.

Ibrahim, I.A., Khatib, T., 2017. A novel hybrid model for hourly global solar radiation prediction using random forests technique and firefly algorithm. Energy Convers. Manag. 138, 413-425.

Janitza, S., Tutz, G., Boulesteix, A.-L., 2016. Random forest for ordinal responses: prediction and variable selection. Comput. Statist. Data Anal. 96, 57-73.

Khlosi, M., Alhamdoosh, M., Douaik, A., Gabriels, D., Cornelis, W., 2016. Enhanced pedotransfer functions with support vector machines to predict water retention of calcareous soil. Eur. J. Soil Sci. 67, 276-284.

Khodaverdiloo, H., Homaee, M., van Genuchten, M.T., Dashtaki, S.G., 2011. Deriving and validating pedotransfer functions for some calcareous soils. J. Hydrol. 399, 93-99.

Klute, A., 1986. Water Retention: Laboratory Methods, In: Klute, A. (Ed.), Methods of Soil Analysis: Part 1—Physical and Mineralogical Methods. Soil Science Society of America, American Society of Agronomy, Madison, WI, pp. 635-662.

Klute, A., Dirksen, C., 1986. Hydraulic Conductivity and Diffusivity: Laboratory Methods, In: Klute, A. (Ed.), Methods of Soil Analysis: Part 1—Physical and Mineralogical Methods. Soil Science Society of America, American Society of Agronomy, Madison, WI, pp. 687-734.

Koekkoek, E., Booltink, H., 1999. Neural network models to predict soil water retention. Eur. J. Soil Sci. 50, 489-495.

Lamorski, K., Pachepsky, Y., Sławiński, C., Walczak, R., 2008. Using support vector machines to develop pedotransfer functions for water retention of soils in Poland. Soil Sci. Soc. Am. J. 72, 1243-1247.

Lamorski, K., Sławiński, C., Moreno, F., Barna, G., Skierucha, W., Arrue, J.L., 2014. Modelling soil water retention using support vector machines with genetic algorithm optimisation. Sci. World J. 2014, 740521, 1-10.

Liaw, A., Wiener, M., 2002. Classification and regression by random forest. R news 2, 18-22.

Ließ, M., Glaser, B., Huwe, B., 2012. Uncertainty in the spatial prediction of soil texture: comparison of regression tree and Random Forest models. Geoderma 170, 70-79.

Liu, Y., 2014. Random forest algorithm in big data environment. Comput. Model. New Tech. 18, 147-151.

Ma, Y., Cukic, B., Singh, H., 2005. A classification approach to multi-biometric score fusion, International Conference on Audio-and Video-Based Biometric Person Authentication. Springer, pp. 484-493.

MathWorks, 2018. MATLAB: the language of technical computing, Inc., Natick, Massachusetts, United States.

Matin, S., Chelgani, S.C., 2016. Estimation of coal gross calorific value based on various analyses by random forest method. Fuel 177, 274-278.

Medrado, E., Lima, J.E., 2014. Development of pedotransfer functions for estimating water retention curve for tropical soils of the Brazilian savanna. Geoderma Regional 1, 59-66.

Merdun, H., Çınar, Ö., Meral, R., Apan, M., 2006. Comparison of artificial neural network and regression pedotransfer functions for prediction of soil water retention and saturated hydraulic conductivity. Soil Tillage Res. 90, 108-116.

Minasny, B., McBratney, A.B., Bristow, K.L., 1999. Comparison of different approaches to the development of pedotransfer functions for water-retention curves. Geoderma 93, 225-253.

Mualem, Y., 1976. A new model for predicting the hydraulic conductivity of unsaturated porous media. Water Resour. Res. 12, 513-522.

Nemes, A., Rawls, W.J., Pachepsky, Y.A., 2005. Influence of organic matter on the estimation of saturated hydraulic conductivity. Soil Sci. Soc. Am. J. 69, 1330-1337.

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

Nemes, A., Schaap, M., Wösten, J., 2003. Functional evaluation of pedotransfer functions derived from different scales of data collection. Soil Sci. Soc. Am. J. 67, 1093-1102.

Neyshaburi, M.R., Bayat, H., Mohammadi, K., Nariman-Zadeh, N., Irannejad, M., 2015. Improvement in estimation of soil water retention using fractal parameters and multiobjective group method of data handling. Arch. Agron. Soil Sci. 61, 257-273.

Nguyen, P.M., Haghverdi, A., De Pue, J., Botula, Y.-D., Le, K.V., Waegeman, W., Cornelis, W.M., 2017. Comparison of statistical regression and data-mining techniques in estimating soil water retention of tropical delta soils. Biosyst. Eng. 153, 12-27.

Pachepsky, Y., Rawls, W., Gimenez, D., Watt, J., 1998. Use of soil penetration resistance and group method of data handling to improve soil water retention estimates. Soil Tillage Res. 49, 117-126.

Pachepsky, Y.A., Rawls, W., 1999. Accuracy and reliability of pedotransfer functions as affected by grouping soils. Soil Sci. Soc. Am. J. 63, 1748-1757.

Pachepsky, Y.A., Rawls, W., Lin, H., 2006. Hydropedology and pedotransfer functions. Geoderma 131, 308-316.

Pachepsky, Y.A., Timlin, D., Varallyay, G., 1996. Artificial neural networks to estimate soil water retention from easily measurable data. Soil Sci. Soc. Am. J. 60, 727-733.

Rab, M., Chandra, S., Fisher, P., Robinson, N., Kitching, M., Aumann, C., Imhof, M., 2011. Modelling and prediction of soil water contents at field capacity and permanent wilting point of dryland cropping soils. Soil Res. 49, 389-407.

Rajkai, K., Kabos, S., Van Genuchten, M.T., 2004. Estimating the water retention curve from soil properties: comparison of linear, nonlinear and concomitant variable methods. Soil Tillage Res. 79, 145-152.

Rawls, W., Brakensiek, D., Soni, B., 1983. Agricultural management effects on soil water processes part I: Soil water retention and Green and Ampt infiltration parameters. Transactions of the ASAE 26, 1747-1752.

Rawls, W., Gish, T., Brakensiek, D., 1991. Estimating soil water retention from soil physical properties and characteristics, Advances in soil science. Springer, pp. 213-234.

Rawls, W.J., Brakensiek, D., 1985. Prediction of soil water properties for hydrologic modeling, Watershed management in the eighties. ASCE, pp. 293-299.

Schaap, M.G., Leij, F.J., van Genuchten, M.T., 1998. Neural network analysis for hierarchical prediction of soil hydraulic properties. Soil Sci. Soc. Am. J. 62, 847-855.

Schaap, M.G., Leij, F.J., van Genuchten, M.T., 2001. Rosetta: A computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. J. Hydrol. 251, 163-176.

Scheinost, A., Sinowski, W., Auerswald, K., 1997. Regionalization of soil water retention curves in a highly variable soilscape, I. Developing a new pedotransfer function. Geoderma 78, 129-143.

Seo, S., 2006. A review and comparison of methods for detecting outliers in univariate data sets, Thesis for Master of Science in Field of Public Health University of Pittsburgh, pp. 1-59.

Shirazi, M.A., Boersma, L., 1984. A unifying quantitative analysis of soil texture. Soil Sci. Soc. Am. J. 48, 142-147.

Shwetha, P., Varija, K., 2015. Soil water retention curve from saturated hydraulic conductivity for sandy loam and loamy sand textured soils. Aquat. Procedia 4, 1142-1149.

Simmons, J.P., Nelson, L.D., Simonsohn, U., 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol. Sci. 22, 1359-1366.

Szabó, B., Szatmári, G., Takács, K., Laborczi, A., Makó, A., Rajkai, K., Pásztor, L., 2019. Mapping soil hydraulic properties using random forest based pedotransfer functions and geostatistics. Hydrol. Earth Syst. Sci. 23, 2615-2635.

Tietje, O., Tapkenhinrichs, M., 1993. Evaluation of pedo-transfer functions. Soil Sci. Soc. Am. J. 57, 1088-1095.

Tomasella, J., Hodnett, M.G., 1998. Estimating soil water retention characteristics from limited data in Brazilian Amazonia. Soil Sci. 163, 190-202.

Tomasella, J., Hodnett, M.G., Rossato, L., 2000. Pedotransfer functions for the estimation of soil water retention in Brazilian soils. Soil Sci. Soc. Am. J. 64, 327-338.

Tóth, B., Makó, A., Toth, G., 2014. Role of soil properties in water retention characteristics of main Hungarian soil types. J. Cent. Eur. Agric. 15, 137-153.

Touil, S., Degre, A., Chabaca, M.N., 2016. Sensitivity analysis of point and parametric pedotransfer functions for estimating water retention of soils in Algeria. Soil 2, 647.

Twarakavi, N.K., Šimůnek, J., Schaap, M., 2009. Development of pedotransfer functions for estimation of soil hydraulic parameters using support vector machines. Soil Sci. Soc. Am. J. 73, 1443-1452.

Ungaro, F., Calzolari, C., Busoni, E., 2005. Development of pedotransfer functions using a group method of data handling for the soil of the Pianura Padano–Veneta region of North Italy: water retention properties. Geoderma 124, 293-317.

van Genuchten, M.T., 1980. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. Soil Sci. Soc. Am. J. 44, 892-898.

Verhagen, J., 1997. Site specific fertiliser application for potato production and effects on N-leaching using dynamic simulation modelling. Agric. Ecosyst. Environ. 66, 165-175.

Verikas, A., Gelzinis, A., Bacauskiene, M., 2011. Mining data with random forests: A survey and results of new tests. Pattern Recognit. 44, 330-349.

Walkley, A., Black, I.A., 1934. An examination of the Degtjareff method for determining soil organic matter, and a proposed modification of the chromic acid titration method. Soil Sci. 37, 29-38.

Wassar, F., Gandolfi, C., Rienzner, M., Chiaradia, E.A., Bernardoni, E., 2016. Predicted and measured soil retention curve parameters in Lombardy region north of Italy. International Soil and Water Conservation Research 4, 207-214.

Wiesmeier, M., Barthold, F., Blank, B., Kögel-Knabner, I., 2011. Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. Plant Soil 340, 7-24.

Williams, J., Ross, P., Bristow, K.L., 1992. Prediction of the Campbell water retention function from texture, structure, and organic matter. In 'Indirect methods for estimating the hydraulic properties of unsaturated soils.' University of California: Riverside.

Wösten, J., Pachepsky, Y.A., Rawls, W., 2001. Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic characteristics. J. Hydrol. 251, 123-150.

Zaklouta, F., Stanciulescu, B., 2012. Real-time traffic-sign recognition using tree classifiers. IEEE Transactions on Intelligent Transportation Systems 13, 1507-1514.

Zhao, P., Su, X., Ge, T., Fan, J., 2016. Propensity score and proximity matching using random forest. Contemp. Clin. Trials 47, 85-92.

Figure captions

~~**Fig. 1.** Correlation matrix plot between input and output variables.~~

~~** Correlation is significant at the *P*<0.01 level.~~

~~* Correlation is significant at the *P*<0.05 level.~~

~~A list of abbreviations is available in the notation box.~~

Fig ~~2~~1. Input variables of the 15 pedotransfer functions (PTFs) for predicting the van Genuchten model parameters (θ_r, θ_s, α and n) of the soil water retention curve (SWRC). A list of abbreviations is available in the notation box.

Fig. ~~3~~2. An architecture of a random forest.

Fig. ~~4~~3. Variation of soil texture classes for the dataset (n = 223) on the United States Department of Agriculture (USDA) textural triangle.

Fig. ~~1~~4. Correlation matrix plot between input and output variables.

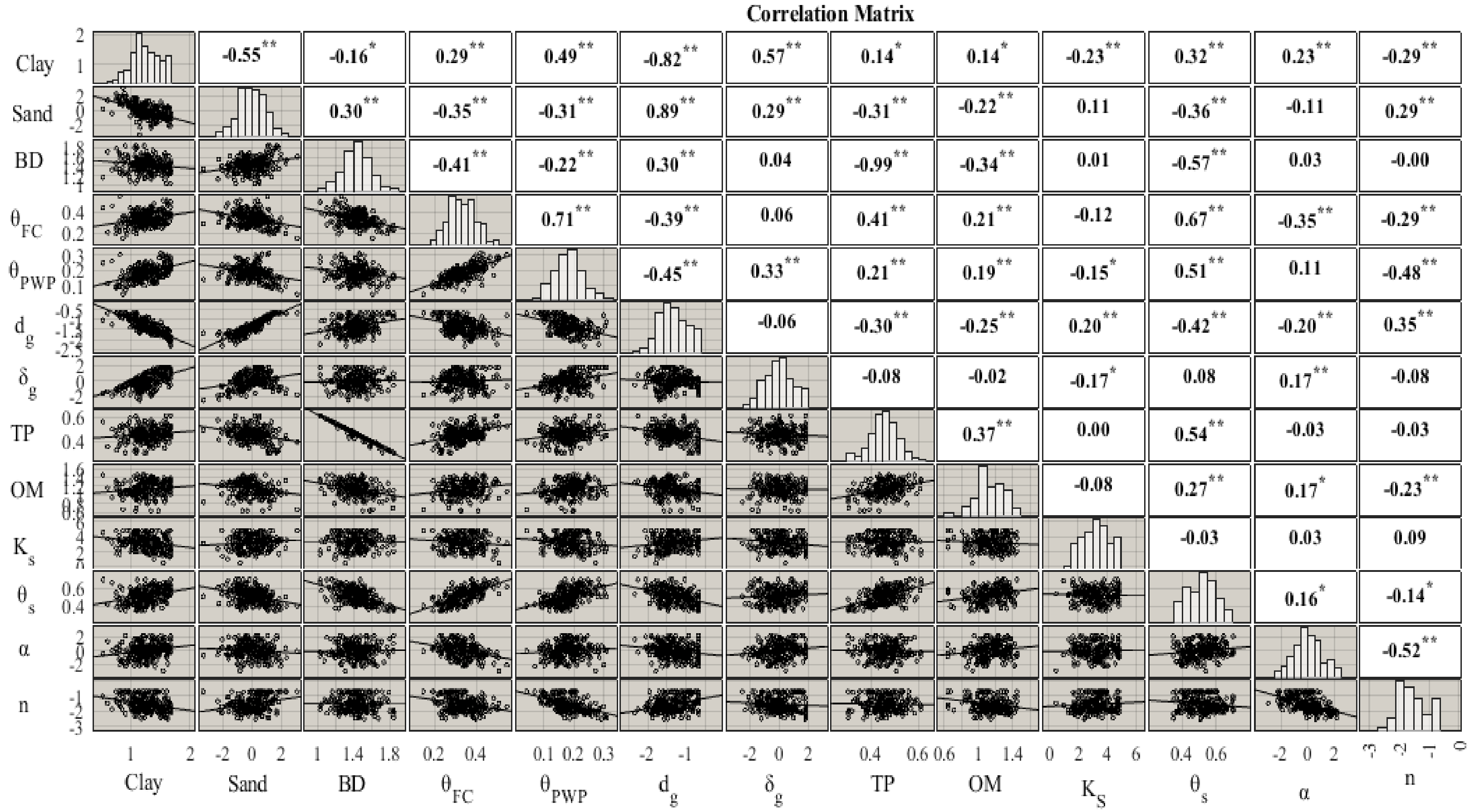
** Correlation is significant at the *P*<0.01 level.

* Correlation is significant at the *P*<0.05 level.

A list of abbreviations is available in the notation box.

Fig. 5. Results of the prediction of the soil water retention curve (SWRC) through the van Genuchten model by the non-linear regression (NLR) and random forests (RF) techniques for the training step as reflected in the integral mean error (*IME*), integral root mean square error (*IRMSE*), coefficient of determination (R_2), and Akaike's information criterion (*AIC*). Vertical lines indicate the standard deviations. Means with the same letter are not significantly different at the significance level of *P*<0.05 (*IRMSE* only).

Fig. 6. Results of the prediction of the soil water retention curve (SWRC) through the van Genuchten model by the Rosetta software, non-linear regression (NLR) and random forests (RF) techniques for the testing step as reflected in the integral mean error (*IME*), integral root mean square error (*IRMSE*), coefficient of determination (R_2), and Akaike's information criterion (*AIC*). Vertical lines indicate the standard deviations. Means with the same letter are not significantly different at the significance level of $P<0.05$ (*IRMSE* only).



~~Fig. 1. Correlation matrix plot between input and output variables.~~

~~** Correlation is significant at the 0.01 level.~~

~~* Correlation is significant at the 0.05 level.~~

~~A list of abbreviations is available in the notation box.~~

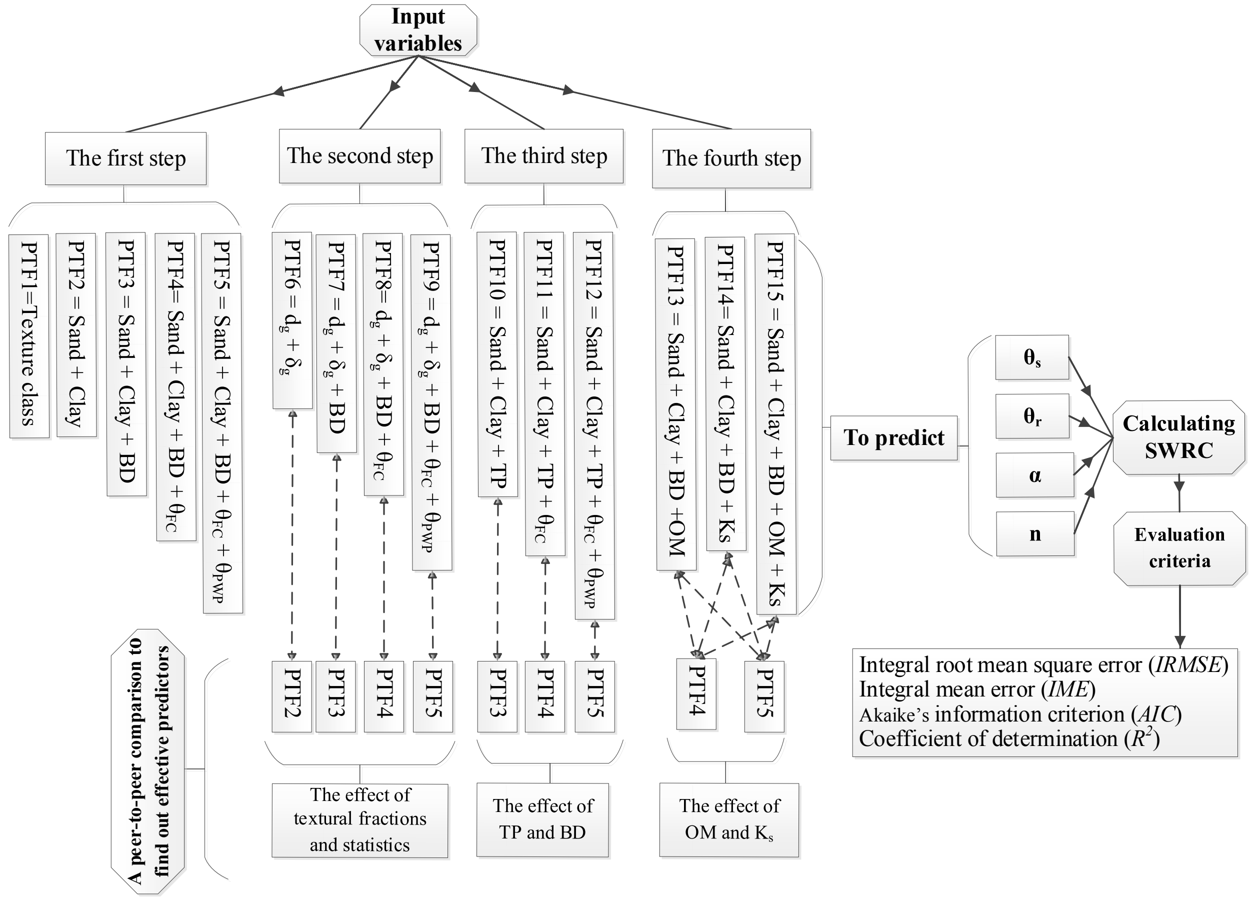

Input variables
The first step
The second step
The third step
The fourth step
PTF1=Texture class
PTF2 = Sand + Clay
PTF3 = Sand + Clay + BD
PTF4= Sand + Clay + BD + θFC
PTF5 = Sand + Clay + BD + θFC + θPWP
PTF6 = dg + δg
PTF7 = dg + δg + BD
PTF8= dg + δg + BD + θFC
PTF9 = dg + δg + BD + θFC + θPWP
PTF10 = Sand + Clay + TP
PTF11 = Sand + Clay + TP + θFC
PTF12 = Sand + Clay + TP + θFC + θPWP
PTF13 = Sand + Clay + BD +OM
PTF14= Sand + Clay + BD + Ks
PTF15 = Sand + Clay + BD + OM + Ks
To predict
θs
θr
α
n
Calculating SWRC
Evaluation criteria
Integral root mean square error (IRMSE)
Integral mean error (IME)
Akaike's information criterion (AIC)
Coefficient of determination (R2)
PTF2
PTF3
PTF4
PTF5
PTF3
PTF4
PTF5
PTF4
PTF5
A peer-to-peer comparison to find out effective predictors
The effect of textural fractions and statistics
The effect of TP and BD
The effect of OM and Ks

Fig ~~2~~1. Input variables of the 15 pedotransfer functions (PTFs) for predicting the van Genuchten model parameters (θ_r, θ_s, α and n) of the soil water retention curve (SWRC). A list of abbreviations is available in the notation box.

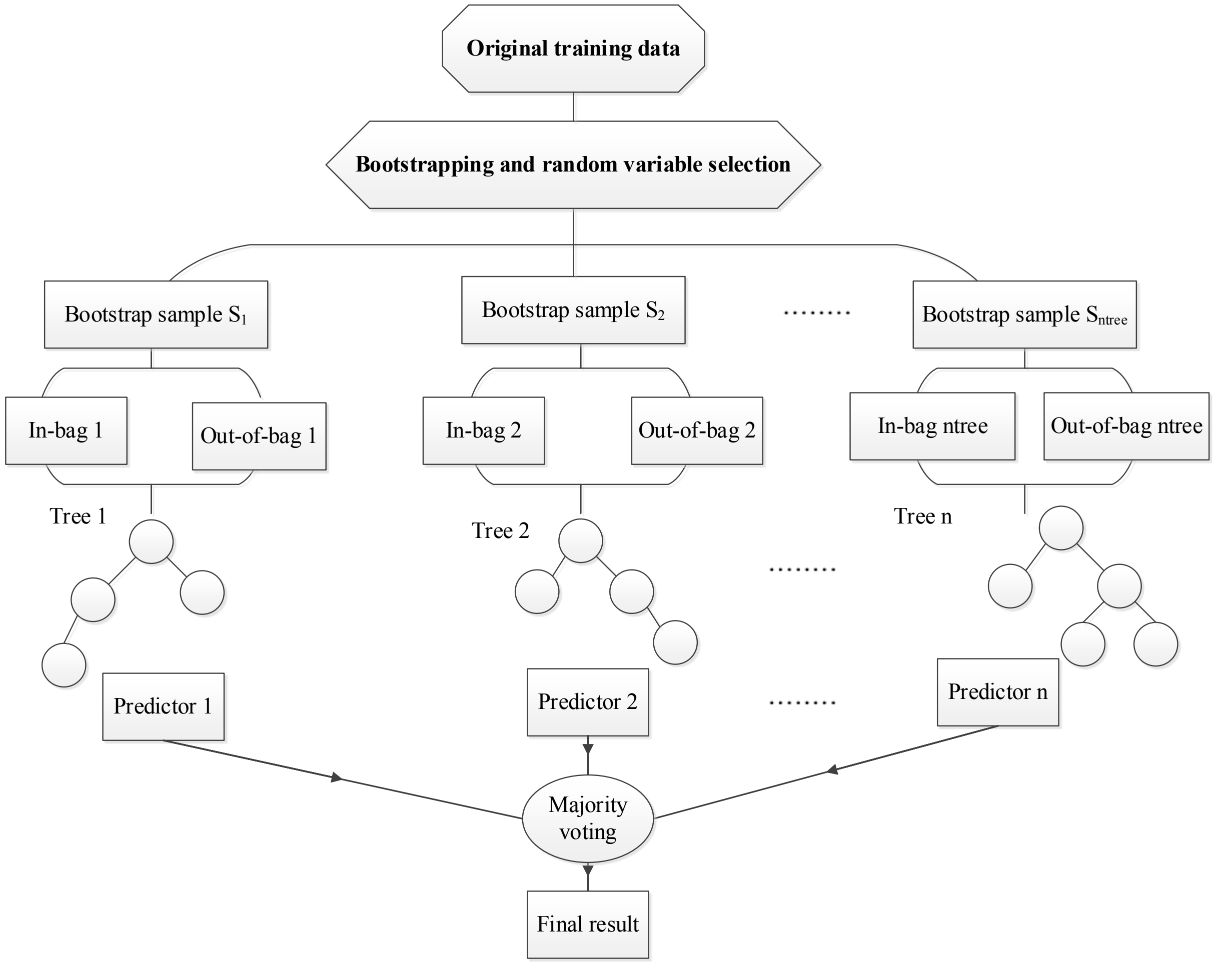


Fig. ~~3~~2. An architecture of a random forest.

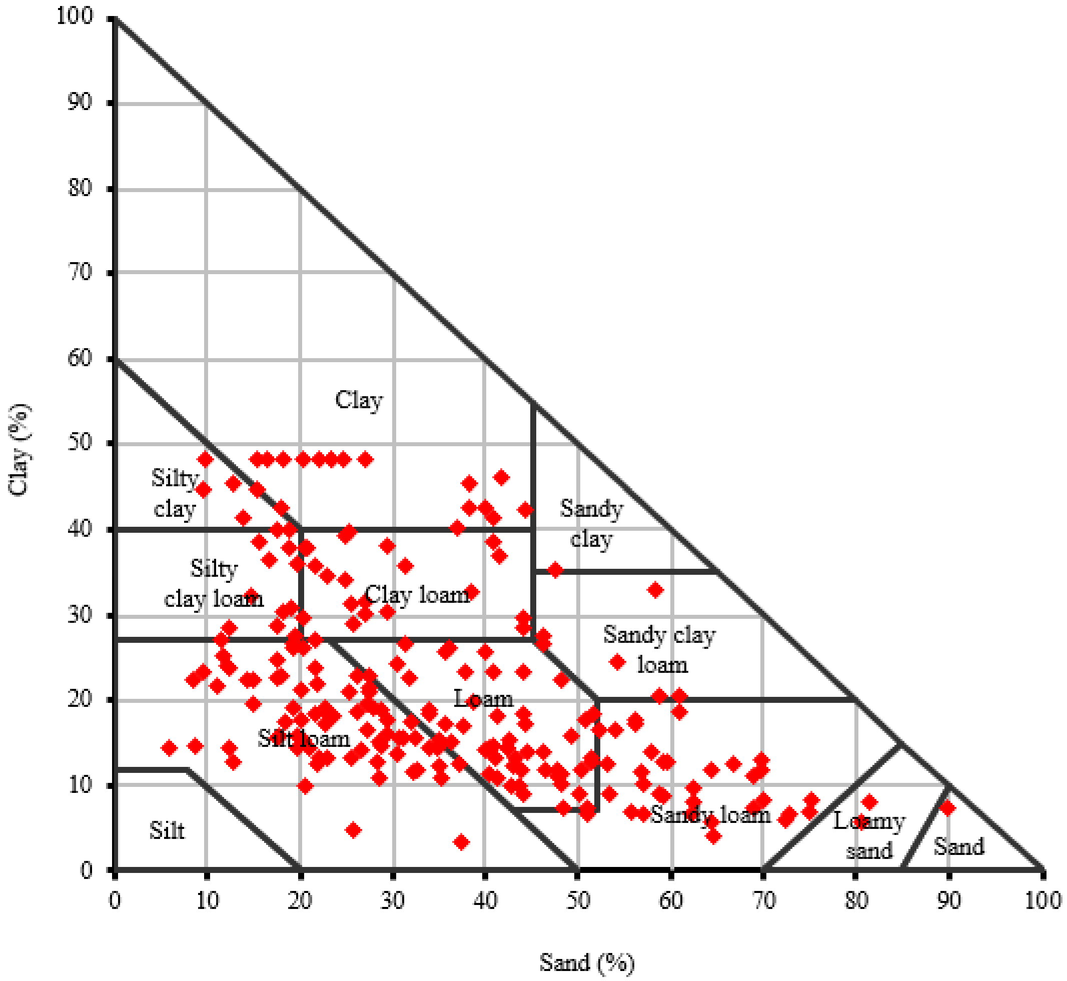


Fig. 43. Variation of soil texture classes for the dataset (n = 223) on the United States Department of Agriculture (USDA) textural triangle.

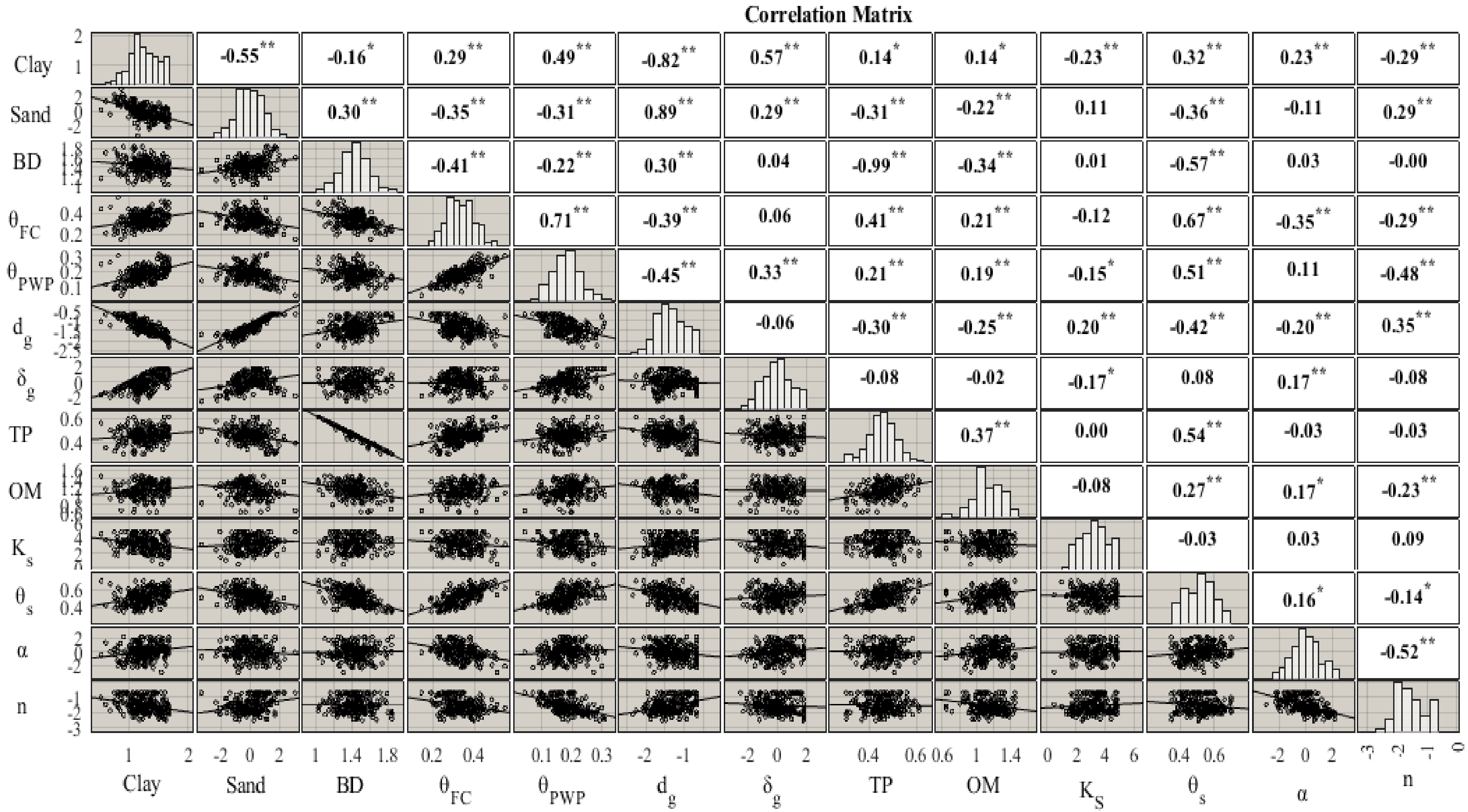
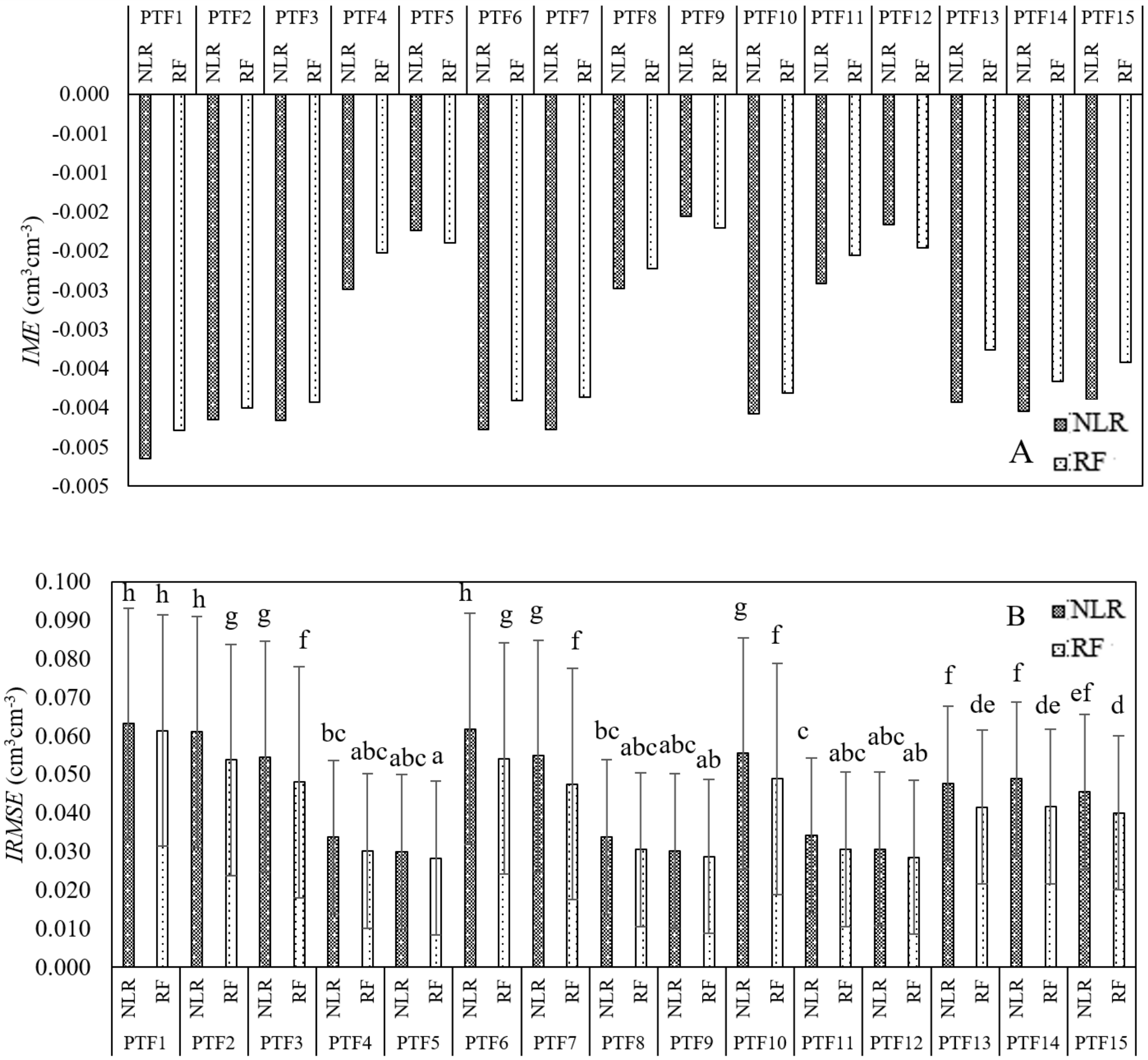


Fig. ~~1~~4. Correlation matrix plot between input and output variables.

** Correlation is significant at the *P*<0.01 level.

* Correlation is significant at the *P*<0.05 level.

A list of abbreviations is available in the notation box.


PTF1
PTF2
PTF3
PTF4
PTF5
PTF6
PTF7
PTF8
PTF9
PTF10
PTF11
PTF12
PTF13
PTF14
PTF15
NLR
RF
IME (cm3cm-3)
0.000
-0.001
-0.001
-0.002
-0.002
-0.003
-0.003
-0.004
-0.004
-0.005
-0.005
A
NLR
RF
IRMSE (cm3cm-3)
0.100
0.090
0.080
0.070
0.060
0.050
0.040
0.030
0.020
0.010
0.000
B
NLR
RF
h
h
h
g
g
f
bc
abc
abc
a
h
g
g
f
bc
abc
abc
ab
g
f
c
abc
abc
ab
f
de
f
de
ef
d

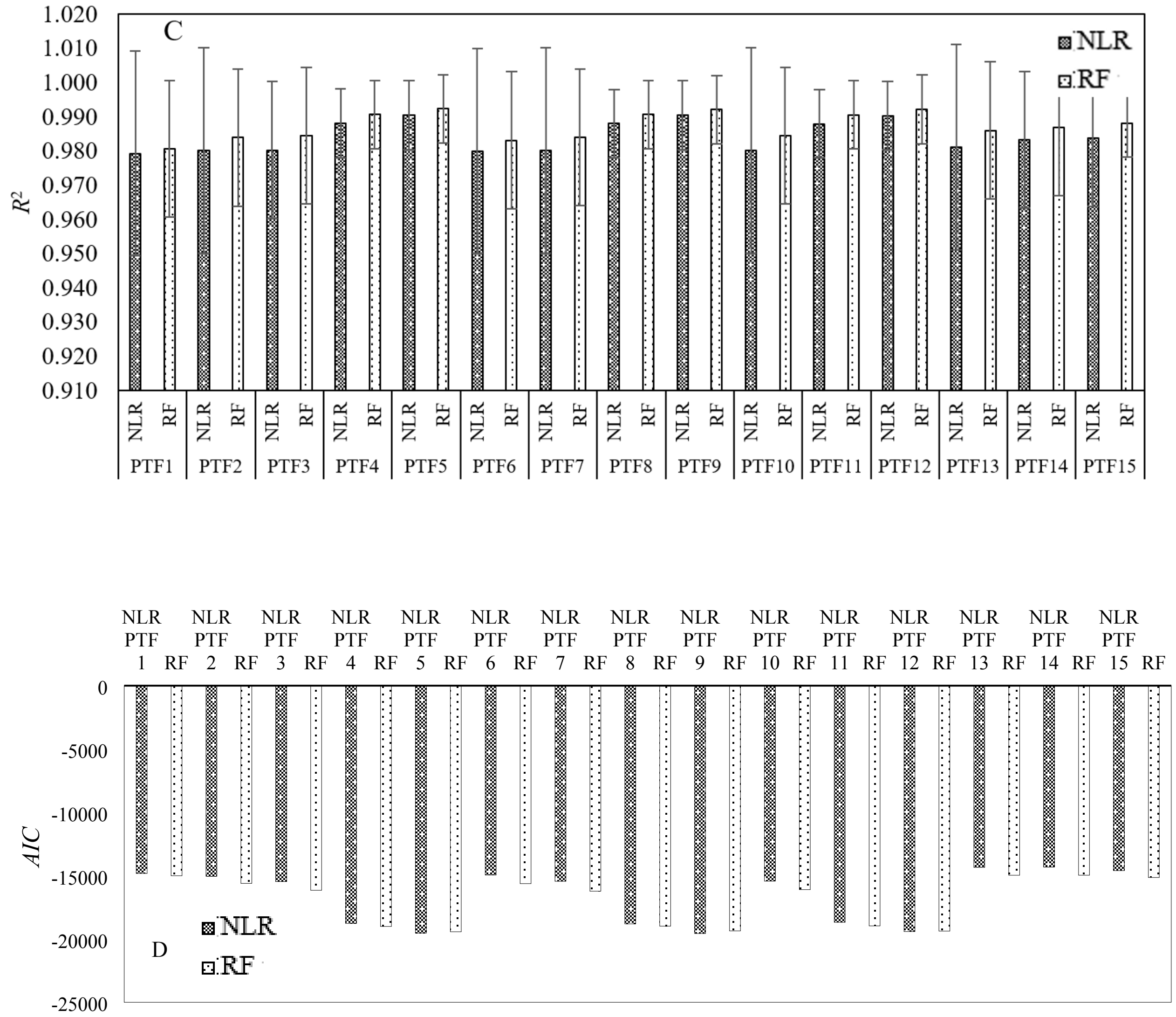
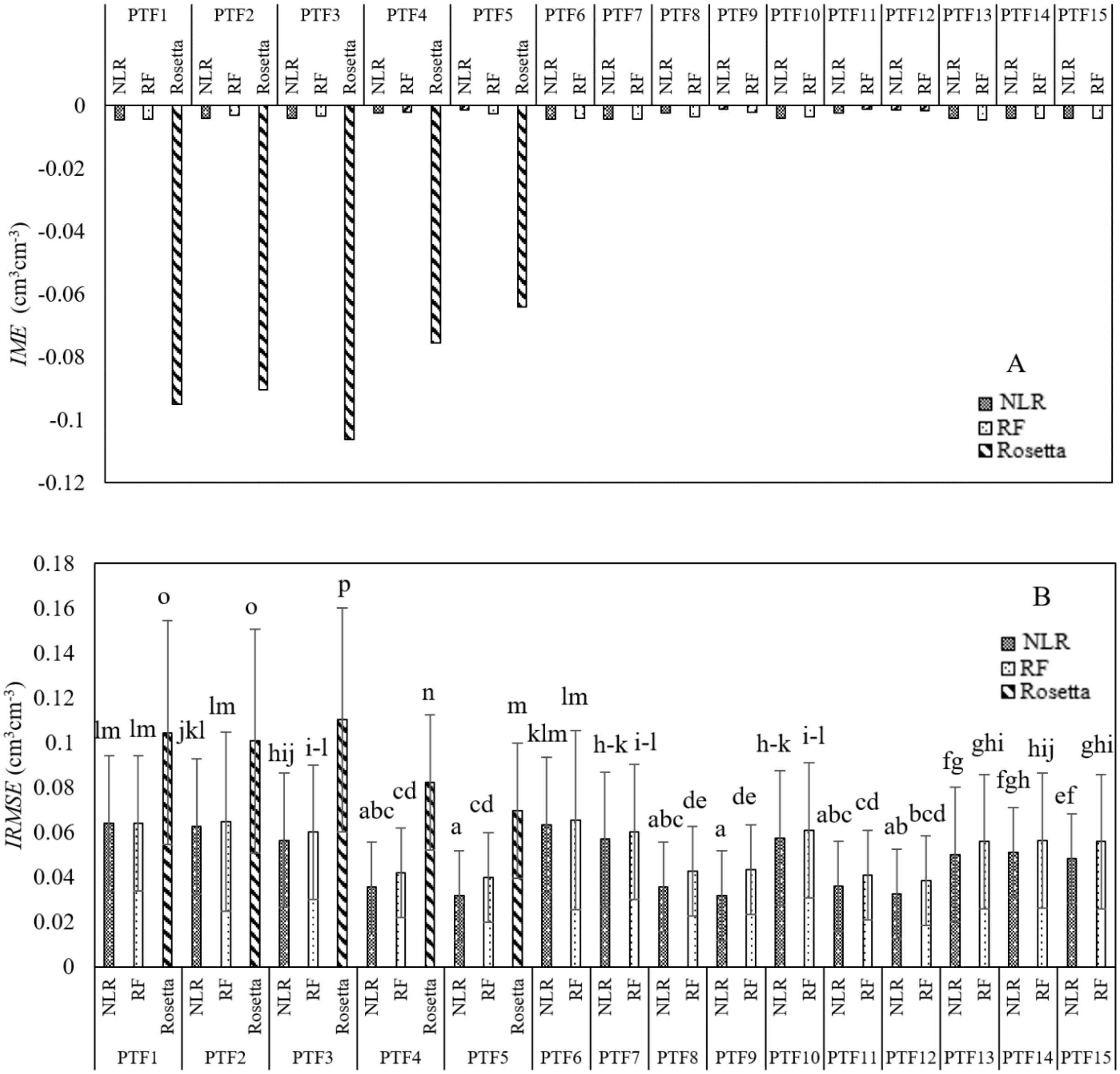


Fig. 5. Results of the prediction of the soil water retention curve (SWRC) through the van Genuchten model by the non-linear regression (NLR) and random forests (RF) techniques for the training step as reflected in the integral mean error (*IME*), integral root mean square error (*IRMSE*), coefficient of determination (*R_2*), and Akaike's information criterion (*AIC*). Vertical lines indicate the standard deviations. Means with the same letter are not significantly different at the significance level of *P*<0.05 (*IRMSE* only).


PTF1
PTF2
PTF3
PTF4
PTF5
PTF6
PTF7
PTF8
PTF9
PTF10
PTF11
PTF12
PTF13
PTF14
PTF15
NLR
RF
Rosetta
0
-0.02
-0.04
-0.06
-0.08
-0.1
-0.12
IME (cm³cm⁻³)
A
▩ NLR
⊡ RF
◪ Rosetta
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
IRMSE (cm³cm⁻³)
B
lm
lm
o
jkl
lm
o
hij
i-l
p
abc
cd
n
a
cd
m
klm
lm
h-k
i-l
abc
de
a
de
h-k
i-l
abc
cd
ab
bcd
fg
ghi
fgh
hij
ef
ghi

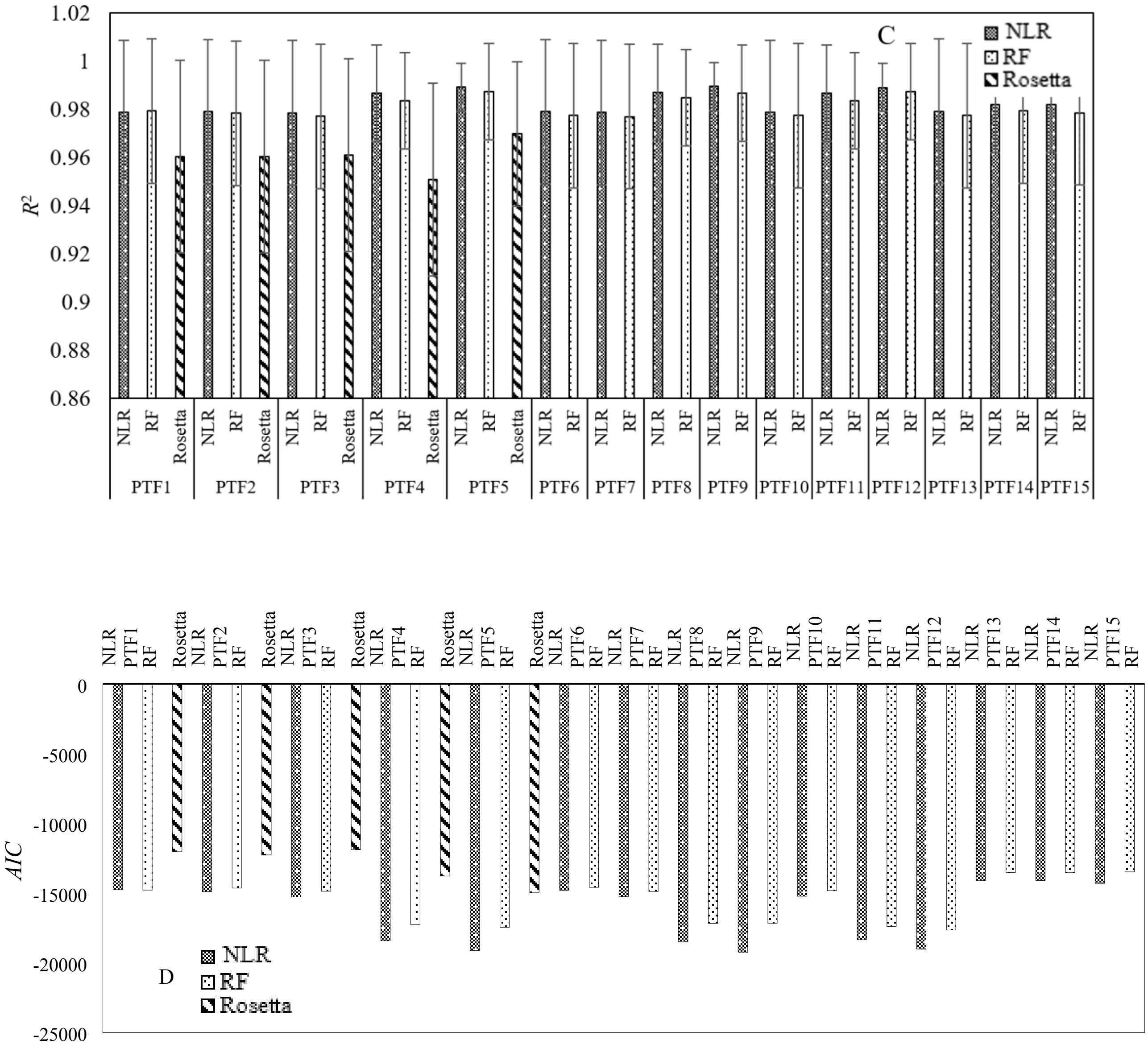


Fig. 6. Results of the prediction of the soil water retention curve (SWRC) through the van Genuchten model by the Rosetta software, non-linear regression (NLR) and random forests (RF) techniques for the testing step as reflected in the integral mean error (*IME*), integral root mean square error (*IRMSE*), coefficient of determination (R_2), and Akaike's information criterion (*AIC*). Vertical lines indicate the standard deviations. Means with the same letter are not significantly different at the significance level of $P<0.05$ (*IRMSE* only).

Table 1- The results of 10, 15 and 20-fold cross-validation (k) for van Genuchten model parameters of the soil water retention curve derived from nonlinear regression (NLR) and random forest (RF) techniques based on root mean square error (*RMSE*) for pedotransfer functions PTF 3, 5 and 11 in the train and test datasets.

			θ_r			θ_s			α			n		
			RMSE			*RMSE*			*RMSE*			*RMSE*		
			Train	Test	Mean	Train	Test	Mean	Train	Test	Mean	Train	Test	Mean
PTF3	k=10	NLR	0.058	0.060	0.059	0.063	0.065	0.064	1.017	1.037	1.027	0.426	0.436	0.431
		RF	0.052	0.061	0.056	0.058	0.073	0.066	0.893	1.084	0.989	0.374	0.442	0.408
	k=15	NLR	0.058	0.060	0.059	0.064	0.064	0.064	1.017	1.030	1.024	0.426	0.434	0.430
		RF	0.052	0.061	0.057	0.058	0.070	0.064	0.894	1.033	0.964	0.374	0.441	0.408
	k=20	NLR	0.058	0.060	0.059	0.064	0.064	0.064	0.064	0.064	0.064	0.426	0.437	0.432
		RF	0.051	0.060	0.056	0.057	0.071	0.064	0.057	0.071	0.064	0.368	0.442	0.405
PTF5	k=10	NLR	0.051	0.053	0.052	0.053	0.054	0.054	0.764	0.796	0.780	0.380	0.397	0.389
		RF	0.043	0.056	0.050	0.046	0.056	0.051	0.675	0.869	0.772	0.327	0.411	0.369
	k=15	NLR	0.051	0.053	0.052	0.053	0.055	0.054	0.764	0.790	0.777	0.381	0.399	0.390
		RF	0.044	0.054	0.049	0.046	0.055	0.050	0.679	0.848	0.763	0.329	0.421	0.375
	k=20	NLR	0.051	0.053	0.052	0.053	0.055	0.054	0.765	0.789	0.777	0.381	0.399	0.390
		RF	0.042	0.054	0.048	0.044	0.054	0.049	0.654	0.842	0.748	0.316	0.412	0.364
PTF11	k=10	NLR	0.058	0.061	0.060	0.065	0.067	0.066	1.018	1.052	1.035	0.431	0.448	0.440
		RF	0.050	0.061	0.056	0.047	0.057	0.052	0.770	0.978	0.874	0.370	0.443	0.406
	k=15	NLR	0.058	0.061	0.060	0.065	0.067	0.066	1.019	1.037	1.028	0.432	0.447	0.439
		RF	0.050	0.060	0.055	0.047	0.057	0.052	0.770	1.009	0.889	0.369	0.450	0.410
	k=20	NLR	0.058	0.060	0.059	0.065	0.065	0.065	1.020	1.024	1.022	0.432	0.439	0.435
		RF	0.049	0.061	0.055	0.046	0.056	0.051	0.745	0.964	0.855	0.361	0.443	0.402

Table ~~1~~2- Some descriptive statistics of the measured soil variables and parameters of the van Genuchten model of the soil water retention curve for the entire dataset (223 soil samples).

Variables[a]	Mean	CV (%)	Minimum	Maximum	*P*-value
Clay content (%)	21.39	54.05	3.47	48.00	0.00
Log (clay content)	1.27	19.08	0.54	1.68	0.66
Sand content (%)	35.45	48.40	5.90	89.80	0.00
Sand content*	-0.01	-14350.94	-3.40	3.14	0.90
Bulk density (g cm^{-3})	1.43	10.97	1.03	1.84	0.83
θ_{FC} (cm^3 cm^{-3})$^{\$}$	0.33	20.44	0.15	0.55	0.45
θ_{PWP} (cm^3 cm^{-3})	0.18	26.21	0.04	0.31	0.90
d_g (mm)	0.07	86.62	0.00	0.21	0.00
Log (d_g)	-1.33	-27.91	-2.34	-0.67	0.77
δ_g (-)	11.57	29.39	4.54	19.97	0.00
δ_g^{*}	-0.01	-9872.87	-2.53	1.80	0.96
Total porosity (cm^3 cm^{-3})	0.46	13.26	0.31	0.61	0.67
Organic matter content (%)	1.84	53.68	0.17	4.41	0.00
(Organic matter content)$^{(1/4)}$	1.13	14.83	0.64	1.45	0.86
K_s (cm day^{-1})	169.10	96.58	0.06	530	0.00
$(K_s)^{(1/4)}$	3.23	30.37	0.50	4.80	0.59
θ_r (cm^3 cm^{-3})	0.04	158.05	0.00	0.17	0.00
θ_s (cm^3 cm^{-3})	0.52	16.26	0.35	0.70	0.56
α (kPa^{-1})	0.06	115.62	0.00	0.29	0.00
α*	0.01	8889.14	-2.93	2.19	0.93
n	1.24	9.80	1.08	1.48	0.00
Ln (n-1)	-1.55	-30.92	-2.52	-0.74	0.05

[a] CV, coefficient of variation.

$^{\$}$. A list of abbreviations is available in the notation box.

* Normalized form of sand content: 0.91+1.06×Ln((sand content- 4.3)/(100.2-sand content)); normalized form of δ_g: -1.04657+1.39359×Asinh((δ_g- 8.4)/3.04); and normalized form of α: 3.6+0.92×Ln((α- 8.2×10^{-6})/(1.6-α)). *P*-value is a significance value for normality test.

Table ~~2~~3- The variance inflation factor (*VIF*) values for normalized form of the input variables.

PTFs	Clay* (%)	Sand (%)	BD$ (g cm^{-3})	θ_{FC} (cm^3 cm^{-3})	θ_{PWP} (cm^3 cm^{-3})	d_g (mm)	δ_g (-)	TP (cm^3 cm^{-3})	OM (%)	K_s (cm day^{-1})
PTF2	1.42	1.42								
PTF3	1.43	1.52	1.10							
PTF4	1.45	1.56	1.25	1.31						
PTF5	1.79	1.58	1.27	2.48	2.56					
PTF6						1.00	1.00			
PTF7			1.11			1.11	1.01			
PTF8			1.25	1.33		1.01	1.22			
PTF9			1.28	2.50	2.73	1.34	1.22			
PTF10	1.55	1.43						1.11		
PTF11	1.58	1.46		1.32				1.26		
PTF12	1.60	1.79		2.49	2.56			1.28		
PTF13	1.48	1.65	1.25						1.14	
PTF14	1.55	1.64	1.14							1.06
PTF15	1.55	1.65	1.25						1.15	1.06

* Normalized form of the input variables is available in Table 2.

$. A list of abbreviations is available in the notation box.

Table ~~3~~4- Analysis of variance of the integral root mean square error (*IRMSE*) of the prediction of the soil water retention curve~~SWRC~~ by different methods (nonlinear regression and random forest) and pedotransfer functions (PTFs 1-15) for both the train and test datasets.

	Source	Degree freedom	Mean square	*F*-value	*P*-value
Train	Repeat (Block)	222	0.007	19.09	<0.0001
	PTFs	14	0.062	180.68	<0.0001
	Methods	1	0.038	109.69	<0.0001
	PTFs × Methods	14	0.001	1.78	0.0356
	Error	6288	0.0003		
Test	Repeat (Block)	222	0.010	16.04	<0.0001
	PTFs	14	0.073	117.22	<0.0001
	Methods	2	0.656	1056.43	<0.0001
	PTFs × Methods	18	0.002	3.68	<0.0001
	Error	7398	0.0006		

- The RF was compared to NLR method and Rosetta-based PTFs to predict the SWRC
- The NLR method had better performance due to higher reliability in the testing step
- The RF and NLR-based PTFs performed better than the Rosetta-based PTFs
- In the absence of moisture points, OM and K_s can be suitable predictors for SWRC
- d_g and δ_g can be suitable alternatives for textural fractions in predicting SWRC
- Total porosity and bulk density have the same effect in predicting the SWRC

Estimating the soil water retention curve: comparison of multiple nonlinear regression approach and random forest data mining technique

M. Rastgou[1], H. Bayat[2]*, and M. Mansoorizadeh[3], Andrew S. Gregory[4]

[1] **Mostafa Rastgou:** Ph. D. Student of Soil Science, Department of Soil Science, Faculty of Agriculture, Bu Ali Sina University, Hamedan, Iran. E-mail: mostafa.rastgo@gmail.com,

[2] **Hossein Bayat:** Associate Professor (Ph. D.), Department of Soil Science, Faculty of Agriculture, Bu Ali Sina University, Hamedan, Iran. Postal Address: Department of Soil Science, Faculty of Agriculture, Bu Ali Sina University, Hamedan, Iran. E-mail: h.bayat@basu.ac.ir Other e-mail: hbayat2001@gmail.com. Office phone: +98-81-34424189, Mobile phone: +98-918-8188378. Fax: +98-81-34424189.

[3] **Muharram Mansoorizadeh:** Assistant professor (Ph. D.), Department of Computer Science, Faculty of Engineering, Bu Ali Sina University, Hamedan, Iran. E-mail: mansoorm@basu.ac.ir

[4] **Andrew S. Gregory**: Sustainable Agriculture Sciences Department, Rothamsted Research, Harpenden, Hertfordshire, AL5 2JQ, UK. E-mail: andy.gregory@rothamsted.ac.uk

*Corresponding author (h.bayat@basu.ac.ir, Other e-mail: hbayat2001@gmail.com**).**

Estimating the soil water retention curve: comparison of multiple nonlinear regression approach and random forest data mining technique

Abstract

This study evaluates the performance of the random forest (RF) method on the prediction of the soil water retention curve (SWRC) and compares its performance with those of nonlinear regression (NLR) and Rosetta-based pedotransfer functions (PTFs), which has not been reported so far. Fifteen RF and NLR-based PTFs were constructed using readily-available soil properties for 223 soil samples from Iran. The general performance of RF and NLR-based PTFs was quantified by the integral root mean square error (*IRMSE*), Akaike's information criterion (*AIC*) and coefficient of determination (R^2). The results showed that the accuracy of the RF-based PTFs was significantly ($P<0.05$) better than the NLR-based PTFs, and that the reliability of the NLR-based PTFs was significantly ($P<0.01$) better than the RF-based PTFs and all of the Rosetta-based PTFs. The average values of the *IRMSE*, *AIC* and R^2 of the RF method were 0.041 cm^3 cm^{-3}, -16997.7, and 0.987, and 0.053 cm^3 cm^{-3}, -15547.5, and 0.981 for the training and testing steps of all PTFs, respectively, whereas the values for the NLR method were 0.046 cm^3 cm^{-3}, -16616.4, and 0.984, and 0.048 cm^3 cm^{-3}, -16355.6, and 0.983 for the training and testing steps, respectively. The PTF5 of the RF and NLR methods, with inputs of sand and clay contents, bulk density, and the water content at field capacity and permanent wilting point, had the greatest R^2 values (0.987 and 0.989, respectively), and the lowest *IRMSE* values (0.039 and 0.032 cm^3 cm^{-3}, respectively) compared to other PTFs for the testing step. Overall, the RF method had less reliability for the prediction of the SWRC compared to the NLR method due to overprediction, uncertainty of determination of forest scale and instability in the testing step. These findings could provide the scientific basis for further research on the RF method.

Keywords: pedotransfer functions; soil water retention curve; soil texture; soil structure; van Genuchten.

Notation

Sand content (%)	S
Clay content (%)	C
Geometric mean diameter (mm)	d_g
Geometric standard deviation (-)	δ_g
Bulk density (g cm^{-3})	BD
Total porosity (cm^3 cm^{-3})	TP
Water content at field capacity, 33 kPa (cm^3 cm^{-3})	θ_{FC}
Water content at 1500 kPa (cm^3 cm^{-3})	θ_{PWP}
Organic matter content (%)	OM
Saturated hydraulic conductivity (cm day^{-1})	K_s
Saturated water content (cm^3 cm^{-3})	θ_s
Residual water content (cm^3 cm^{-3})	θ_r
Random forest	RF
Nonlinear regression	NLR
Soil water retention curve	SWRC

1 Introduction

Soil hydraulic properties are principle factors that control the movement of water and solutes in the soil. Determination of the soil hydraulic properties is required for many distinct applications linked with irrigation, land use planning, drainage and drought risk assessment (Dobarco et al., 2019). The soil water retention curve (SWRC) is one of the most important soil hydraulic properties. It defines the relationship between soil matric potential and soil water content (Hillel, 1998). The SWRC is a crucial parameter in soil and water management for sustainable and improved agricultural production (Shwetha and Varija, 2015). The SWRC depends principally on texture, structure and bulk density (BD) of soils (Wassar et al., 2016). Many methods have been introduced for the direct measurement of the SWRC in the laboratory (e.g., the hanging water column and pressure plate methods) (Klute, 1986) and in the field (e.g., tensiometric) (Bruce and Luxmoore, 1986). Measurements of the SWRC at several matric potentials can be expensive, difficult and time-consuming, hence it is common to predict it by modelling (Dobarco et al., 2019). Modelling of soil water is an essential tool in evaluating the effects of different managements on crop yield and environmental quality (Verhagen, 1997).

Pedotransfer functions (PTFs) translate easy-to-measure data that we have (e.g., texture class, particle size distribution (PSD) and BD) into difficult-to-measure data that we need (soil hydraulic data) (Bouma, 1989). Estimates of the SWRC by PTFs are valuable in many studies, such as hydrology, soil mapping and hydrogeology (Børgesen and Schaap, 2005). The point- and parametric-based PTFs are generally developed to predict water content at specific matric potential values and the entire SWRC, respectively, by multiple linear (MLR) and nonlinear regression (NLR) methods (Gunarathna et al., 2019b; Merdun et al., 2006; Minasny et al., 1999; Rajkai et al., 2004; Tomasella et al., 2000). Data mining techniques including artificial neural networks (ANNs) (Bayat et al., 2013a; Bayat et al., 2013b; Gunarathna et al., 2019a; Koekkoek and Booltink, 1999; Pachepsky et al., 1996), group method of data handling (GMDH) (Bayat et al., 2011; Neyshaburi et al., 2015; Pachepsky and Rawls, 1999), nonparametric nearest neighbor technique (Botula et al., 2013; Gunarathna et al., 2019a; Haghverdi et al., 2015; Nemes et al., 2006; Nguyen et al., 2017) and support vector machine (SVM) (Khlosi et al., 2016; Lamorski et al., 2008; Lamorski et al., 2014; Twarakavi et al., 2009), have been applied successfully for PTF development.

Random forest (RF), or random decision forests, has become a popular approach as an ensemble learning method for prediction and classification (Verikas et al., 2011). The RF method has been developed by Breiman (2001) as an expansion of the classification and regression trees (CART) technique to provide better performance of prediction results (Wiesmeier et al., 2011). So far, few studies have been carried out on the application of the RF method in soil science. Tóth et al. (2014) applied the RF method to analyze the relationship between soil water content at four matric suctions (0.1, 33, and 1500 kPa, and 150 MPa) and Hungarian soil map information. They found that the importance of soil properties in the prediction of the soil water content varied according to soil type and matric suction. Recently Szabó et al. (2019) have developed PTFs based on RF and geostatistics methods to map soil hydraulic properties, such as water contents at saturation, field capacity and wilting point, for the Balaton catchment area in Hungary. Araya and Ghezzehei (2019) compared the performances of four machine-learning algorithms including the k-nearest neighbors (kNNs), support vector regression (SVR), RF, and boosted regression tree (BRT) for prediction of saturated hydraulic conductivity. They found that the BRT model outperformed the other algorithms closely followed by the RF model. Gunarathna et al. (2019a) tested three machine-learning algorithms including ANN, kNN, and RF to estimate volumetric water content at matric suctions of 10, 33 and 1500 kPa for soils in Sri Lanka. They recommended that the PTFs to be developed using the RF algorithm. Ließ et al. (2012) studied uncertainty in the spatial prediction of soil texture by comparison of the RF and regression tree techniques for 56 soil profiles and found that the former method provided a better result. Also, Wiesmeier et al. (2011) utilized the RF technique to develop digital mapping of the soil organic matter content in 120 soil profiles. They found that the prediction accuracy of the RF modeling was acceptable. A review of literatures therefore revealed that the RF data mining technique has been applied to develop PTFs to predict specific points of the SWRC, such as field capacity and permanent wilting point, or particular properties such as saturated hydraulic conductivity, but it has not been used to develop parametric-based PTFs of the van Genuchten model parameters, so far. Therefore, the objective of the present study was to develop simple parametric-PTFs to predict the SWRC with greater accuracy and reliability using a novel approach with the RF data mining technique. We compare its performance with those of the multiple NLR approach and with Rosetta software (Schaap et al., 2001) on the prediction of the SWRC through finding the best input variables and PTFs for the SWRC.

2 Materials and methods

2.1 Sample collection and determination

In the present study 223 undisturbed and disturbed soil samples were taken from six provinces of Iran including west Azarbaijan (35° 8□ − 39° 46□ N, 44° 3□ − 47° 23□ E; 60 data), Hamedan (33° 59□ − 35° 48□ N, 47° 34□ − 49° 36□ E; 55 data), Kermanshah (33° 41□ − 35° 17□ N, 45° 24□ − 48° 6□ E; 26 data), Kurdistan (34° 45□ − 36° 31□ N, 45° 31□ − 48° 13□ E; 22 data), Mazandaran (35° 46□ − 36° 58□ N, 50° 21□ − 58° 08□ E; 30 data) and Fars (27° 2□ − 31° 42□ N, 50° 42□ − 55° 38□ E; 30 data). Steel cylinders, measuring 5.1 cm in diameter and 3.5 cm in height, were used to collect the undisturbed samples. Since the sampling was done from different locations of the various provinces, the topsoil and subsoil layers of soil at different locations had different depths and thicknesses. We collected samples from the center of the topsoil and subsoil layers, which represented the pedological A and B horizons, respectively. The sampling depths varied from 10 to 35 cm for topsoil (208 samples) and from 20 to 45 cm for subsoil (15 samples), reflecting the variation in the soil profiles.

Soil PSD was analyzed by the hydrometer method (Gee and Or, 2002), and the geometric mean and standard deviation of particle diameter (d_g and δ_g, respectively) were calculated by equations from Shirazi and Boersma (1984). Organic matter (OM) content was determined by the Walkley and Black (1934) method and BD by the core method (Blake and Hartge, 1986). Total porosity (TP) was calculated from BD and particle density, and the saturated hydraulic conductivity (K_s) was measured with a constant head permeameter (Klute and Dirksen, 1986). The SWRC was constructed by measuring the volumetric water content at matric suctions of 0 (saturation status of soil samples), 1, 2 and 5 kPa with a sandbox apparatus, and at 10, 25, 50, 100, 200, 500, 1000 and 1500 kPa with a pressure plate apparatus. Undisturbed samples were used for measurement of the matric suctions from 0 to 100 kPa and disturbed samples were used for matric suctions from 200 to 1500 kPa. Two key points in the SWRC are the water contents at field capacity (30 kPa suction; θ_{FC}) and permanent wilting point (1500 kPa suction; θ_{PWP}).

2.2 Soil-water retention equation

The van Genuchten–Mualem (Eq. (1)) model (Mualem, 1976; van Genuchten, 1980) was utilized to describe the SWRC data.

$$\theta = \theta_r + \left(\theta_s - \theta_r\right) \times \frac{1}{\left[1+\left(\alpha h\right)^n\right]^{\left(1-\frac{1}{n}\right)}} \tag{1}$$

where θ_r and θ_s are residual and saturated water contents (cm^3 cm^{-3}), respectively, and h is the soil water suction (kPa). The parameter α is related to the inverse of the air entry pressure (>0, kPa^{-1}) and n (>1, dimensionless parameter) is related to the pore size distribution of the soil (van Genuchten, 1980). In the present study, van Genuchten model parameters θ_r, θ_s, α and n were obtained using the MATLAB software (MathWorks, 2018).

2.3 Data pre-processing

Data pre-processing and regression assumptions, including detection of outliers, normality test of the residuals, multicollinearity and independence of the residuals, were applied for all variables (Berry, 1993). The outliers in the data were identified by the inter-quartile range (IQR) method (Seo, 2006) and were replaced by the lower and upper threshold values (MathWorks, 2018). Before developing PTFs, all variables were evaluated by Kolmogorov-Smirnov normality and multicollinearity tests by the SPSS 24 software (IBM, 2016). The degree of multicollinearity in the PTFs was tested by the variance inflation factor (VIF=1/1-R^2_j, where R^2_j is the R^2 value obtained by regressing the j^{th} predictor on the remaining predictors) (Hocking, 2013). Also, to avoid multicollinearity between textural contents, the silt fraction was not used as a predictor. The variables clay content, sand content, d_g, δ_g, OM, K_s, α and n had non-normal distributions, therefore, transformations were applied to normalize them.

2.4 Developing PTFs

The PTF inputs were arranged in four steps (Fig. 1). The first step (PTFs 1-5) was based on basic soil properties (i.e., sand content (%), clay content (%), BD (g cm^{-3}), θ_{FC} (cm^3 cm^{-3}) and θ_{PWP} (cm^3 cm^{-3})) according to Rosetta-based PTFs (Schaap et al., 2001) for comparison of SWRC estimates by other methods. The parameters of the van Genuchten model were predicted in all steps. In the second step (PTFs 6-9), d_g (mm) and δ_g were used as new inputs instead of sand and clay contents in the previous step to evaluate the efficiency of using statistical descriptors of PSD to predict the parameters of the van Genuchten model. To build the third step (PTFs 10-12), TP (cm^3 cm^{-3}) replaced BD from PTFs 3-5 to evaluate the effect of using TP instead of BD on the prediction of the parameters of the van Genuchten model. In other words, the purpose of the second and third steps was to evaluate whether the use of another form of descriptors of soil structure (TP instead of the BD) and soil texture (d_g and δ_g instead of the sand and clay contents) would improve the accuracy of the estimates or not. In the last step, PTFs 13-15 were developed by including OM (%) and K_s (cm day^{-1}) as new variables to evaluate the efficiency of these instead of the water content at specific matric suctions on the prediction of the van Genuchten model parameters. The input variables of the 15 PTFs are shown in Fig. 1.

To compare the results of PTFs 1-5 of the RF and NLR methods with those of the Rosetta models, the parameters of the van Genuchten model (θ_r, θ_s, α and n) were estimated by the PTFs built in the Rosetta software (PTFs 1-5), using the measured values of input variables based on PTFs 1-5 as predictors in the Rosetta program. The estimated coefficients of the van Genuchten model were used to calculate the estimated water content at matric suctions from 0 to 1500 kPa (estimated SWRCs). Then curve-by-curve comparison of the measured and estimated SWRCs was performed with different evaluation statistics. Since there is no training step in the Rosetta software, the results of the Rosetta model was only compared with the results of the testing step.

To evaluate the effect of using different descriptors of PSD on the prediction of the SWRC, PTFs 6, 7, 8 and 9 from the second step were compared with PTFs 2, 3, 4 and 5 from the first step, respectively (Fig. 1). In the same way, to evaluate effect of using different descriptors of soil structure on the prediction of the SWRC, PTFs 10, 11 and 12 from the third step were compared with PTFs 3, 4 and 5 from the first step, respectively. Also, the PTFs 13-15 were compared with the PTFs 4 and 5 to find out the efficiency of OM and K_s variables as predictors (Fig. 1).

Fig 1.

In the present study, the k-fold cross validation approach (Efron and Tibshirani, 1994) was utilized to obtain training and testing datasets for each PTF. The number of folds (i. e., k) was obtained by trial and error. To do so, some PTFs, selected randomly, were developed with 10, 15 and 20-fold cross-validation. Then, the k value which resulted in the best performance of the PTFs, was selected to develop all PTFs in this study. The results showed that 20-fold cross validation performed better than the other folds in most of the PTFs (Table 1). Therefore, 20-fold cross validation was selected to develop PTFs in this study. Based on this approach, the 223 samples were randomly divided into 20 subsets and 20 models were developed by each predicting technique for each PTF. In each model, training and testing datasets were based on a ratio of 19:1. Finally, the average of the results of 20 models was calculated for each PTF. Therefore, all data were used for the training and testing steps of the PTFs.

Table 1-

2.5 Description of modeling techniques

2.5.1 Multiple nonlinear regression

A NLR model based on a second-order polynomial for the prediction of the response variable *y* from a number of *p* predictors can be written as (Rawls and Brakensiek, 1985):

$$y = a + \sum_{i=1}^{p} \left(b_i x_i + c_i x_i^2 \right) \qquad (2)$$

where *a* is the intercept, and two regression coefficients b_i and c_i are determined for every input variable x_i.

2.5.2 Random forest: an ensemble of regression trees

RF has become a popular tool for regression and classification problems. The RF is an ensemble method based on the regression tree methodology (i.e., CART) that was introduced for better performance (Breiman, 2001). The model building process in the RF is the same as that in the CART method but without pruning (Breiman, 1984). Also, whereas a regression tree only grows by a single tree the RF grows by forest of trees. In other words, unlike a regression tree, in the RF for each tree only a subset of the input variables is applied. The number of inputs in each tree and also the number of trees in the forest can be distinct and it depends on the dataset. Least-squares boosting (LSBoost) fits regression ensembles. At every step, the ensemble fits a new learner to the difference between the observed response and the aggregated prediction of all learners grown previously. The ensemble fits to minimize the mean-squared error (MathWorks, 2018). The number of trees used here was 16 which was established by trial and error. An architecture of the RF algorithm is shown in Fig. 2 where input matrix X consists of N samples and M input variables (sample set S = [(x_i, y_i), i = 1, 2, …, N], (X, Y) $\in R^M \times R$). The bootstrap method is utilized to construct *n* tree sample sets from the sample set S. At each bootstrap sample, about one-third of the dataset S was utilized as out of the bootstrap data or out-of-bag (*OOB*) data; whereas the rest is called in-bag data (Ibrahim and Khatib, 2017) (Fig. 2). Modeling of the regression tree is done for each sample set. In the RF algorithm, all individual trees give a predictive result. The final prediction value is calculated based on an average result of all individual trees (Wiesmeier et al., 2011). The prediction error is defined as follows (Liaw and Wiener, 2002):

$$MSE_{OOB} = \frac{\sum_{i=1}^{n_{tree}} \left(y_i - \hat{y}_i^{OOB} \right)^2}{n_{tree}} \tag{3}$$

where MSE_{OOB} is the mean square error of the *OOB* data prediction, n_{tree} is the number of trees, and y_i and $\hat{y}_i^{OOB}$ are the actual value of the *OOB* data and the average of all *OOB* predictions, respectively. Among all the ensemble methods, the RF method has high capability in solving classification and regression problems, because the RF method combines several simple regression trees to better optimize prediction (Zaklouta and Stanciulescu, 2012). The RF method increases differences for each single tree through random selection of the training samples and different variables at each splitting node. In the present study, the NLR and RF algorithms were implemented by fitnlm and fitensemble functions in the MATLAB software, respectively. (MathWorks, 2018).

Fig. 2.

2.6 Evaluation criteria

The estimated water content was computed by estimated parameters of the van Genuchten model for each PTF at matric suctions from 0 to 1500 kPa. For curve-by-curve comparison of the measured and predicted SWRCs, different evaluation statistics were used. Various statistical criteria including integral root mean square error (*IRMSE*), integral mean error (*IME*) (Tietje and Tapkenhinrichs, 1993), Akaike's information criterion (*AIC*) (Akaike, 1974) and coefficient of determination (R^2) (Wösten et al., 2001), were utilized to assess the predictive ability of the RF and NLR algorithms, which are defined as:

$$IRMSE\left(cm^3 cm^{-3}\right)=\left[\frac{1}{b-a}\int_a^b\left(\hat{y}_i-y_i\right)^2 d\log|h|\right]^{\frac{1}{2}} \tag{4}$$

$$IME\left(cm^3 cm^{-3}\right)=\frac{1}{b-a}\int_a^b\left(\hat{y}_i-y_i\right)d\log|h| \tag{5}$$

$$AIC=N\times\ln\left[\sum_{i=1}^{N}\frac{\left(y_i-\hat{y}_i\right)^2}{N}\right]+2P \tag{6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{N}\left(y_i - \bar{y}_i\right)^2} \tag{7}$$

where *h* is the matric suction (kPa), y_i , $\hat{y}_i$ and $\bar{y}_i$ are the measured, predicted and average of the measured values of the water content, respectively, *a* and *b* values define the matric suction range over which the experimental curve is measured, i.e., 0 and 1500 kPa, respectively, and *P* and *N* are the number of parameters and the number of points that were considered in the SWRC, respectively. In calculating the *AIC*, *N* is the total number of points that were considered in the SWRC of all soil samples (i. e., *N*= number of soil samples × number of paired points of the suction-water content for each soil sample), and *i* is paired points of the suctions-water content of the SWRC of each soil sample.

To evaluate the performance of each method in different PTFs, the effect of method as the first factor at two levels in the training step (*i.e.*, NLR and RF methods) and at three levels in the testing step (*i.e.*, NLR, RF and Rosetta methods), and the different PTFs as the second factor at 15 levels (PTF1 to PTF15), were investigated using a two-way analysis of variance (ANOVA) with a randomized complete block design, based on the *IRMSE* of prediction of the SWRC. The *IRMSE* criterion calculates the total error, including bias and random errors, and is a more appropriate criterion for evaluating the accuracy and reliability of the RF and NLR methods compared to other criteria (Chai and Draxler, 2014). Therefore, to compare the predicting accuracy and reliability of the RF and NLR methods, the average values of the *IRMSE* was compared with Duncan's test by MathWorks (2018) software.

3 Results and discussion

3.1 *Descriptive statistics of the soil properties*

Table 2 summarizes some basic descriptive statistics for soil variables of the entire dataset used for the development of the PTFs. It can be seen that the average and maximum of clay content were 21.4 and 48%, respectively. The OM ranged from 0.17 to 4.41% with a mean of 1.84%, which was low due to the arid and semi-arid climates of Iran. The variation in soil texture is shown graphically in the United States Department of Agriculture (USDA) textural triangle (Fig. 3). Considering the distribution and range of the variables (Fig. 3 and Table 2), the dataset can be considered as representative of soils in arid and semi-arid regions of Iran.

Table 2

Fig. 3.

3.2 *Correlation of input and output variables*

The simple correlation coefficients between all variables are depicted by matrix plot in Fig. 4. Correlation analysis was done between normalized input and output variables. The correlation test was not performed for the θ_r variable, because its value was zero in 138 out of 223 soil samples, as has been reported in other studies (Campbell and Horton Jr, 2002; Rawls et al., 1991; Tomasella et al., 2000) for θ_r variable. Clay and sand contents, θ_{FC}, θ_{PWP}, d_g and OM had the greatest significant correlations with the parameters of the van Genuchten model (Fig. 4), which was consistent with other studies (Dexter et al., 2008; Nemes et al., 2006). For example, the correlation coefficient between clay content and θ_s ($r = 0.323$) is close to that between OM and θ_s ($r = 0.268$). Also, the results showed that there were significant correlations between θ_{PWP} and input variables of clay content (+), sand content (–), BD (–), OM (+) and K_s (–), and also between θ_{PWP} and θ_s (+) and n (–) parameters of the van Genuchten model (Fig. 4). Botula et al. (2012) also found the same observation for the correlation of θ_{PWP} with sand and clay contents and BD of tropical Lower Congo soils. Nevertheless, with regard to these correlation coefficients, clay and sand contents, θ_{FC}, d_g and OM can be used for developing PTFs to estimate the SWRC. On the contrary, there was no correlation between K_s and the van Genuchten model parameters. There are many cases, where two variables might not show a strong simple correlation, but may show a strong association in the regression, along with other predictors. In other words, the simple correlation coefficient is a way to show the relationship between independent and dependent variables, but it cannot show a model for the relationship between these two variables, when other independent variables have been used in a multiple regression (Simmons et al., 2011). The result of multiple regression analysis with backward selection method showed that the K_s variable remained in the PTF14 and PTF15 for all the van Genuchten model parameters. Some of the regression equations with backward selection method are shown in the following as examples:

$$\theta_r=-0.69+0.22\times\text{Clay}+0.278\times\text{Sand}+0.20\times\text{K}_\text{s},\ R=0.31^{**} \quad (8)$$

$$\alpha=-3.72+0.23\times\text{Clay}+0.17\times\text{BD}+0.282\times\text{K}_\text{s},\ R=0.33^{**} \quad (9)$$

$$n=-1.76+0.24\times\text{Sand}+0.164\times\text{K}_\text{s},\ R=0.30^{**} \quad (10)$$

On the other hand, the non-linear correlations between variables are very important in this study. Both the multiple NLR approach and RF data mining technique are non-linear prediction methods. Fig. 4 only shows simple linear correlation between variables, but there may be non-linear correlations between variables, which may affect the estimation of the dependent variables. For example, the results of non-linear correlations showed that K_s had strong correlations with θ_s and α of the van Genuchten model parameters by logarithmic (θ_s=0.652-0.027×lnK_s, R=0.62**) and power (α=0.007×$K_s^{0.283}$, R=0.57**) equations, respectively, which were greater than their simple correlations

Fig. 4.

3.3 Development of the PTFs using the RF and NLR methods

Results of the multicollinearity analysis (*VIF)* are shown in Table 3. The *VIF* values showed low levels of multicollinearity among the independent variables (*VIF*<10) (Khodaverdiloo et al., 2011).

Table 3-

3.3.1 Comparing the accuracy and reliability of the RF and NLR methods

Table 4 shows the results of the ANOVA of the *IRMSE* of prediction of the SWRC by different methods and PTFs. The effect of methods and PTFs, and their interaction, on the *IRMSE* was significant at P<0.01, 0.01 and 0.05, respectively, in the training step, and at P<0.01, 0.01 and 0.01, respectively, in the testing step. Therefore, we focus on the results and discussion of the comparison of the method × PTF interaction effects.

Table 4-

Results of the prediction of the SWRC through the van Genuchten model using the NLR and RF-based PTFs are depicted in Figs. 5 and 6 for the training and testing steps, respectively. The accuracy and reliability are used to express the performance of the PTFs in the training and testing steps, respectively.

The results of the first to fourth steps of the training dataset (Fig. 5) showed that the RF method had better performance compared to the NLR method for the prediction of the SWRC in all PTFs in terms of the *IRMSE* and R^2 criteria and the differences were significant (P<0.05) for PTFs 2, 3, 6, 7, 10, 13, 14 and 15 in terms of the *IRMSE* criterion. Also, the accuracy of the RF method was better than that of the NLR method in 80% of the PTFs (with the exception of the PTFs 5, 9 and 12) in terms of the *AIC* criterion. In the training step, the values of the *IRMSE* of the first to fourth steps for the NLR model varied from 0.030 to 0.063 cm^3 cm^{-3} and these were larger than those in the RF model, which ranged from 0.028 to 0.061 cm^3 cm^{-3}, respectively. Also, the values of the R^2 of the first to fourth steps for the RF model varied from 0.981 to 0.992, and this was larger than those in the NLR model, which ranged from 0.979 to 0.991 (Fig. 5).

The results of the first to fourth steps of the testing dataset (Fig. 6) showed that the NLR method had a better performance compared to the RF method on the prediction of the SWRC for PTFs 5, 8, 9 and 15 only in terms of the *IRMSE* criterion (significant at $P<0.05$). In the other PTFs there were no significant differences between the *IRMSE* of the two methods and the R^2 and *AIC* criteria were comparable. In the testing step, the values of the *IRMSE* and *AIC* of the first to fourth steps for the RF models varied from 0.038 to 0.065 cm^3 cm^{-3} and from -13476.2 to -17646.8, respectively, and these were comparable to those of the NLR models (with the exception of PTF1), which ranged from 0.032 to 0.064 cm^3 cm^{-3} and from -14096.1 to -19234.1, respectively (Fig. 6). Also, the values of the R^2 of the first to fourth steps for the NLR models varied from 0.979 to 0.989, and this was comparable to those of the RF models for all PTFs, which ranged from 0.977 to 0.987 (Fig. 6).

In each of the PTFs 1 to 5, the NLR and RF methods performed better ($P<0.05$) than the Rosetta PTFs. Fig. 6(A) shows that the Rosetta-based PTFs had greater values of the *IME* criterion compared to the NLR and RF-based PTFs. The reason can be attributed to the various methods of optimizing parameters. The Rosetta method has only one ANN type with particular structure. In other words, the number of hidden layers (one) and neurons (six) and also the activation function (tangent hyperbolic) are constant for prediction of the SWRC in the Rosetta software. Therefore, the Rosetta method is not a dynamic approach for optimization, whereas the parameters of the RF method, such as number of splits and trees, and learning rate continuously and dynamically, change to achieve the best result of the objective function. The Rosetta method was developed from a large dataset, while the soils used in the present study were collected from a completely different climate area that was not represented in the Rosetta's database. Also, presented RF and NLR models were trained using this particular dataset while Rosetta had been trained using a different dataset. In other words, the results of the PTFs in the testing step were based on a soil dataset used for training. This could be a reason for Rosetta's poor performance compared with the RF and NLR methods. As a result, it seems that the universal portability of the Rosetta method can be limited. The testing results are in agreement with Touil et al. (2016) who found that the parametric-based PTFs of nonlinear models gave a better prediction than the Rosetta PTFs. The Figs. 5(A) and 6(A) showed that all of the *IME* values were negative for all PTFs at the training and testing steps. There are regular errors (bias) in the prediction of the SWRC that can be corrected by finding a correction coefficient, which would improve the accuracy and reliability of the estimations (Bayat et al., 2015).

Fig. 5.

Fig. 6.

The RF method in the training section gave better predictions of the SWRC compared to the NLR method (Fig. 5). The RF method produces low bias and variation in the data by majority voting compared to a single regression tree (Cheng et al., 2019; Matin and Chelgani, 2016). In this connection, the results of the standard deviations (SD) of evaluation criteria in each PTF for the training step (Fig. 5) showed that the RF method had a lower variation than the NLR method. Accordingly, the values of SD for the *IRMSE* and R^2 criteria were 0.024 and 0.022, respectively, for the NLR model and these were larger than those in the RF model, which were 0.020 and

0.017, respectively, for the training step. On the other hand, the RF method can be applied to high dimensional datasets in regressions (Janitza et al., 2016; Zhao et al., 2016).

As depicted in Fig. 6, unlike in the training section, the NLR method gave better predictions in the testing section compared to the RF method for the prediction of the SWRC. In other words, the reliability of the NLR method was better than that of the RF method in all the PTFs. The NLR equations can be more useful than the MLR method for the prediction of the SWRC due to their high flexibility (Williams et al., 1992). In other words, the NLR models have capacity to capture nonlinear relationships in the dataset. Tomasella et al. (2000) successfully developed parametric PTFs for soils of the humid tropics using polynomials of n^{th} order. Medrado and Lima (2014) successfully developed NLR-based PTFs to predict the four parameters of the van Genuchten model for Brazilian soils. Also, Touil et al. (2016) developed parametric-PTFs to predict the SWRC using the NLR method from more readily-available properties such as soil texture, OM content, and BD for 242 soil samples of Algeria. They reported that the parametric-PTFs had better performance than Rosetta-based PTFs.

In the present study, in contrast to the NLR method which had less differences between the error values of the training and testing steps, the error values of the RF method in the testing dataset were much greater than those in the training dataset. These results can be due to overprediction phenomenon in the RF method. Gupta et al. (2017) expressed that one of the disadvantages of the RF method is the overprediction. In other words, the RF method is a ‘greedy’ method that easily leads to overprediction and instability in the testing step and solving this problem can be of great significance for improving the reliability of the RF method (Liu, 2014). Also, Ma et al. (2005) reported instability in results of the RF method. The forest size developed by the RF has not been clearly defined (Liu, 2014). Therefore, oversized scale can decrease the reliability and efficiency of the SWRC prediction. Hong et al. (2016) evaluated landslide susceptibility maps produced using the RF method and compared these maps with those from statistical-based methods, such as logistic regression, and their study revealed that the performance of the statistical-based methods was better than that of the RF method. A similar result was reported by Esposito et al. (2014). Generally, RFs are best suited for problems with many input variables and a reasonable sample size. According to our results (Figs. 5 and 6), performance of the PTFs was improved by increasing the number of input variables.

3.3.2 Evaluation of the effect of the basic soil properties on prediction performance of the SWRC

A significant improvement was achieved in the accuracy of PTF5 (with the inputs of Sand content+Clay content+BD+θ_{FC}+θ_{PWP}) compared to other PTFs (with the exception of PTFs 4, 8, 9, 11 and 12) by both NLR and RF methods in terms of the *IRMSE* criterion (Fig. 5). Among the PTFs of each method (RF or NLR), PTF5 had the greatest R^2 (0.992 and 0.991, respectively) and the smallest *IRMSE* (0.028 and 0.03, respectively) and *AIC* (-19432 and -19571.1, respectively) in the training step of the prediction of the SWRC. In connection with the importance of input variables, an improvement was achieved in the reliability of the prediction of the SWRC by PTFs 9 (with the inputs of d_g+δ_g+BD+θ_{FC}+θ_{PWP}) and 12 (with the inputs of Sand content+Clay content+TP+θ_{FC}+ θ_{PWP}) from the second and third steps, using the NLR (*IRMSE*=0.032 cm^3 cm^{-3}, *AIC*=-19234.1 and R^2=0.989) and RF (*IRMSE*=0.038 cm^3 cm^{-3}, *AIC*=-17646.8 and R^2=0.987) methods, respectively, in comparison with the other PTFs of each method (Fig 6). However, the differences of PTFs 9 and 12 were not significant (*$P<0.05$*) with PTFs 4, 5, 8, 11 and 12 in the NLR method and with PTFs 4, 5, 8, 9 and 11 in the RF method, respectively, in terms of the *IRMSE* criterion.

3.3.2.1 Effect of using different input variables of PSD and soil structure as predictors on the SWRC prediction

To evaluate the effect of using different descriptors of the PSD on the prediction of the SWRC, PTFs 2, 3, 4 and 5 (clay and sand contents) from the first step were compared with PTFs 6, 7, 8 and 9 (d_g and δ_g) from the second step, respectively. In the same way, to evaluate the effect of using different descriptors of soil structure on the prediction of the SWRC, PTFs 3, 4 and 5 (BD) were compared with PTFs 10, 11 and 12 (TP) from the third step, respectively. The accuracy and reliability of the prediction of the SWRC by both NLR and RF methods were not significantly different ($P<0.05$) (Figs. 5B and 6B). For descriptors of soil structure, the accuracy and reliability of the prediction of the SWRC by both NLR and RF methods decreased in terms of the *IRMSE* criterion for PTFs 10 to 12 from the third step compared to PTFs 3 to 5 (with the exception of PTFs 11 and 12 in the testing step for the RF method), respectively, when TP was used instead of BD in the list of input variables (Figs. 5B and 6B). However, the differences were not significant ($P<0.05$).

The lack of significant differences between textural contents (clay and sand contents) and statistics (d_g and δ_g), and also between TP and BD on the SWRC prediction can be due to correlation of these parameters with the parameters of the van Genuchten model (Fig. 4). The SWRC is strongly influenced by the soil structure or pore-size distribution and soil texture at small and great matric suctions, respectively (Pachepsky et al., 2006). Therefore, input variables of the textural contents or statistics can influence the residual saturation region of the SWRC. However, soil water content at the dry end (high matric suctions) of the SWRC is primarily determined by textural contents (Hillel, 1998). Also, TP and BD are indicators of soil structure and had significant correlations with θ_s (Fig. 4). Indeed, TP was calculated by BD and particle density (Rab et al., 2011).The d_g and δ_g predictors were derived from soil textural contents (Shirazi and Boersma, 1984). Therefore, these could be reasons for similar effects of textural contents and statistics and also TP and BD predictors on the prediction of the SWRC.

Many researchers used textural contents (Adhikary et al., 2008; Chakraborty et al., 2011; Minasny et al., 1999; Tomasella and Hodnett, 1998), d_g and δ_g (Rab et al., 2011; Scheinost et al., 1997; Ungaro et al., 2005), BD (Bayat et al., 2011; Pachepsky et al., 1998) and TP (Bayat et al., 2011; Pachepsky et al., 1998; Schaap et al., 1998) as effective predictors to derive point- and parametric-PTFs. Nemes et al. (2003), Schaap et al. (2001) and Schaap et al. (1998) reported that the variables of PTF5 have better capability on predicting the parameters of the van Genuchten (1980) model with an average *RMSE* of 0.026, 0.044 and 0.058 cm^3cm^{-3}, respectively.

According to the results of the accuracy (Fig. 5) and reliability (Fig. 6) of PTFs 5, 9 and 12, it seems that certain points of the SWRC (e.g., θ_{FC}) can help to improve the prediction of the SWRC and this is in agreement with Schaap et al. (2001). These results indicate that the presence of at least one moisture point (e.g., θ_{FC}) can improve the prediction of the SWRC. In the first step, PTF5 with two moisture points (θ_{FC}+θ_{PWP}) and PTF4 with one moisture point (θ_{FC}) improved the prediction of the SWRC by 55, 48, 42% and 51, 44, 38% in terms of the *IRMSE* criterion compared to the PTFs 1, 2 and 3, respectively, in the RF method in the training step. In the testing section of the second step, PTF9 with two moisture points (θ_{FC}+θ_{PWP}) and PTF8 with one moisture point (θ_{FC}) decreased the *IRMSE* by 49, 44% and 44, 39% compared to PTFs 6 and 7, respectively, in the NLR method. The points above are also true for the RF-based PTF12 in the third step of the testing section. Many researchers successfully applied θ_{FC} and θ_{PWP} as effective predictors to derive point- and parametric-PTFs (Børgesen and Schaap, 2005; Nemes et al., 2003; Schaap et al., 2001; Touil et al., 2016; Twarakavi et al., 2009).

3.3.2.2 Effect of using OM and K_s as predictors on the SWRC prediction

To evaluate the effect of using OM and/or K_s and points of the SWRC on the prediction of the SWRC, the performances of PTFs 13, 14 and 15 were compared with those of PTFs 4 and 5. The accuracy and reliability of the prediction of the SWRC by both NLR and RF methods, significantly ($P<0.05$) decreased in terms of the *IRMSE*, for the PTFs 13, 14 and 15 from the fourth step, when OM and/or K_s were used with textural contents and BD as inputs instead of θ_{FC} or both θ_{FC} and θ_{PWP} in the list of input variables, compared to PTFs 4 and 5 at the first step (Figs. 5B and 6B). Therefore OM and K_s were not as effective predictors as θ_{FC} and θ_{PWP} in the prediction of the SWRC, because θ_{FC} and θ_{PWP} are two points of the SWRC and enter direct information of the SWRC into the PTFs, whereas OM and K_s enter indirect information, and therefore had less effect in the improvement of the estimation of the SWRC. These results agreed well with results obtained by Børgesen and Schaap (2005). They reported that PTFs with the inputs of θ_{FC} and θ_{PWP} had smaller *RMSE* values than a PTF with the input of OM (0.038 versus 0.042) in the prediction of the SWRC. On the other hand, the results showed that by adding OM and/or K_s as predictors in the PTFs 13, 14 and 15, the accuracy (Fig. 5B) and reliability (Fig. 6B) of the prediction of the SWRC improved by 16, 13, 17 and 7.1, 6.3, 6.9%, respectively, compared to the PTF3 in terms of the *IRMSE* criterion in the RF method.

The SWRC depends mainly on the soil texture and structure (Hillel, 1998), with OM affecting the SWRC through development of soil structure (Nemes et al., 2005), important at low suctions. However, the OM retains water itself. Similarly, K_s can be a descriptive index of soil texture and porosity (Hillel, 1998). The correlation results showed that K_s can be strongly influenced by clay content and textural statistics (d_g and δ_g) (Fig. 4). Bayat et al. (2013b) applied OM and K_s to estimate water content at the measured matric suctions. They found that the OM and K_s can be most appropriately used in point-based PTFs to estimate water content at the matric suctions of 25 and 50 kPa. Also, the result of the present study agreed well with results obtained by Hollis et al. (1977) and Rawls et al. (1983). In this study, the OM and K_s in the PTFs 13, 14 and 15 were not effective predictors compared to θ_{FC} and θ_{PWP} in the PTFs 4 and 5, otherwise they had better results than PTF3.

4 Conclusion

Machine-learning tools have been widely applied for the prediction of the SWRC. The present study evaluated the capability and performance of the RF method as a novel machine learning tool and compared its performance with that of the NLR method on the prediction of the SWRC, using different combinations of easily-available soil properties. It was found that the RF method had a better performance ($P<0.05$) than the NLR method in the training step of the prediction of the SWRC in term of the *IRMSE*, *AIC* and R^2 criteria. However, in the testing step, NLR had a better performance than RF. The poor performance of the RF compared to the NLR method could be due to overprediction in the former, resulting in instability in the testing step. The RF method can be sensitive to sparse areas on the prediction space. In other words, the performance and sensitivity of predictions, and the computational intensity of the RF method depends on the distribution and number of observations and input variables. Therefore, the method should be tested further with different datasets to evaluate its performance through soil and water investigations. An improvement was achieved in the accuracy of the prediction of the SWRC in the training step of the PTF5 (with the inputs of Sand content+Clay content+BD+θ_{FC} +θ_{PWP}) by both NLR and RF methods and also an improvement was achieved in the reliability of the PTF9 (with the inputs of d_g+δ_g+BD+θ_{FC}+θ_{PWP}) and PTF12 (with the inputs of Sand content +Clay content+TP+ θ_{FC}+θ_{PWP}) by the NLR and RF methods compared to other PTFs, respectively. Considering that the PTFs 5, 9, and 12 had no significant difference from PTF4 (with the inputs of Sand content+Clay content+BD+θ_{FC}) and PTF8 (with the inputs of d_g+δ_g+BD+θ_{FC}+θ_{PWP}), these latter PTFs, with less and more-easily measured input variables, are suggested to be the best PTFs for the prediction of the SWRC. Also, PTFs without predictors of θ_{FC} and θ_{PWP}, such as the PTF3 (with the inputs of Sand content+Clay content+BD) and PTF7 (with the inputs of d_g+ δ_g+BD), can be effective models for the prediction of the SWRC.

Acknowledgements

This work was funded by Bu Ali Sina University, Hamedan, Iran. The authors are deeply grateful to anonymous reviewers and the editor for their helpful comments on the manuscript.

References

Adhikary, P.P., Chakraborty, D., Kalra, N., Sachdev, C., Patra, A., Kumar, S., Tomar, R., Chandna, P., Raghav, D., Agrawal, K., 2008. Pedotransfer functions for predicting the hydraulic properties of Indian soils. Soil Res. 46, 476-484.

Akaike, H., 1974. A new look at the statistical model identification. IEEE transactions on automatic control 19, 716-723.

Araya, S.N., Ghezzehei, T.A., 2019. Using Machine Learning for Prediction of Saturated Hydraulic Conductivity and Its Sensitivity to Soil Structural Perturbations. Water Resour. Res. 55, 5715-5737.

Bayat, H., Ersahin, S., Hepper, E.N., 2013a. Improving estimation of specific surface area by artificial neural network ensembles using fractal and particle size distribution curve parameters as predictors. Environ. Model Assess. 18, 605-614.

Bayat, H., Neyshabouri, M., Mohammadi, K., Nariman-Zadeh, N., 2011. Estimating water retention with pedotransfer functions using multi-objective group method of data handling and ANNs. Pedosphere 21, 107-114.

Bayat, H., Neyshaburi, M.R., Mohammadi, K., Nariman-Zadeh, N., Irannejad, M., 2013b. Improving water content estimations using penetration resistance and principal component analysis. Soil Tillage Res. 129, 83-92.

Bayat, H., Sedaghat, A., Sinegani, A.A.S., Gregory, A.S., 2015. Investigating the relationship between unsaturated hydraulic conductivity curve and confined compression curve. J. Hydrol. 522, 353-368.

Berry, W.D., 1993. Understanding regression assumptions. Sage Publications, London.

Blake, G., Hartge, K., 1986. Bulk density, Methods of Soil Analysis: Part 1. Physical and Mineralogical Methods, Madison, Wisconsin, USA: Soil Sci. Soc. Am. J.

Børgesen, C.D., Schaap, M.G., 2005. Point and parameter pedotransfer functions for water retention predictions for Danish soils. Geoderma 127, 154-167.

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

Botula, Y.-D., Cornelis, W.M., Baert, G., Mafuka, P., Van Ranst, E., 2013. Particle size distribution models for soils of the humid tropics. Journal of Soils and Sediments 13, 686-698.

Bouma, J., 1989. Using soil survey data for quantitative land evaluation, Advances in soil science. Springer, pp. 177-213.

Breiman, L., 1984. Classification and regression trees. Routledge, New York.

Breiman, L., 2001. Random forests. Machine learning 45, 5-32.

Bruce, R.R., Luxmoore, R.J., 1986. Water Retention: Field Methods, In: Klute, A. (Ed.), Methods of Soil Analysis: Part 1—Physical and Mineralogical Methods. Soil Science Society of America, American Society of Agronomy, Madison, WI, pp. 663-686.

Campbell, G.S., Horton Jr, R., 2002. Methods of Soil Analysis: Part 4, Physical Methods. Soil Sci. Soc. Am.

Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)?– Arguments against avoiding RMSE in the literature. Geosci. Model Dev. 7, 1247-1250.

Chakraborty, D., Mazumdar, S., Garg, R., Banerjee, S., Santra, P., Singh, R., Tomar, R., 2011. Pedotransfer functions for predicting points on the moisture retention curve of Indian soils. Indian J. Agr. Sci. 81, 1030.

Cheng, L., Chen, X., De Vos, J., Lai, X., Witlox, F., 2019. Applying a random forest method approach to model travel mode choice behavior. Travel behaviour and society 14, 1-10.

Dexter, A., Czyż, E., Richard, G., Reszkowska, A., 2008. A user-friendly water retention function that takes account of the textural and structural pore spaces in soil. Geoderma 143, 243-253.

Dobarco, M.R., Cousin, I., Le Bas, C., Martin, M.P., 2019. Pedotransfer functions for predicting available water capacity in French soils, their applicability domain and associated uncertainty. Geoderma 336, 81-95.

Efron, B., Tibshirani, R.J., 1994. An introduction to the bootstrap. CRC press.

Esposito, C., Barra, A., Evans, S.G., Scarascia Mugnozza, G., Delaney, K., 2014. Landslide susceptibility analysis by the comparison and integration of random forest and logistic regression methods; application to the disaster of Nova Friburgo-Rio de Janeiro, Brasil (January 2011), EGU General Assembly Conference Abstracts.

Gee, G.W., Or, D., 2002. 2.4 Particle-Size Analysis, In: Dane, J.H., Topp, C.G. (Eds.), Methods of Soil Analysis: Part 4 Physical Methods. Soil Science Society of America, Madison, WI, pp. 255-293.

Gunarathna, M., Sakai, K., Nakandakari, T., Momii, K., Kumari, M., 2019a. Machine Learning Approaches to Develop Pedotransfer Functions for Tropical Sri Lankan Soils. Water 11, 1940.

Gunarathna, M., Sakai, K., Nakandakari, T., Momii, K., Kumari, M., Amarasekara, M., 2019b. Pedotransfer functions to estimate hydraulic properties of tropical Sri Lankan soils. Soil Till. Res. 190, 109-119.

Gupta, B., Rawat, A., Jain, A., Arora, A., Dhami, N., 2017. Analysis of various decision tree algorithms for classification in data mining. Int. J. Comput. Appl. 163, 15-19.

Haghverdi, A., Leib, B.G., Cornelis, W.M., 2015. A simple nearest-neighbor technique to predict the soil water retention curve. Transactions of the ASABE 58, 697-705.

Hillel, D., 1998. Environmental soil physics: Fundamentals, applications, and environmental considerations. Academic press.

Hocking, R.R., 2013. Methods and applications of linear models: regression and the analysis of variance. John Wiley & Sons.

Hollis, J., Jones, R., Palmer, R., 1977. The effects of organic matter and particle size on the water-retention properties of some soils in the West Midlands of England. Geoderma 17, 225-238.

Hong, H., Pourghasemi, H.R., Pourtaghi, Z.S., 2016. Landslide susceptibility assessment in Lianhua County (China): a comparison between a random forest data mining technique and bivariate and multivariate statistical models. Geomorphology 259, 105-118.

IBM, C., 2016. IBM SPSS Statistics for Windows, Version 24.0. Armonk, NY: IBM Corp.

Ibrahim, I.A., Khatib, T., 2017. A novel hybrid model for hourly global solar radiation prediction using random forests technique and firefly algorithm. Energy Convers. Manag. 138, 413-425.

Janitza, S., Tutz, G., Boulesteix, A.-L., 2016. Random forest for ordinal responses: prediction and variable selection. Comput. Statist. Data Anal. 96, 57-73.

Khlosi, M., Alhamdoosh, M., Douaik, A., Gabriels, D., Cornelis, W., 2016. Enhanced pedotransfer functions with support vector machines to predict water retention of calcareous soil. Eur. J. Soil Sci. 67, 276-284.

Khodaverdiloo, H., Homaee, M., van Genuchten, M.T., Dashtaki, S.G., 2011. Deriving and validating pedotransfer functions for some calcareous soils. J. Hydrol. 399, 93-99.

Klute, A., 1986. Water Retention: Laboratory Methods, In: Klute, A. (Ed.), Methods of Soil Analysis: Part 1—Physical and Mineralogical Methods. Soil Science Society of America, American Society of Agronomy, Madison, WI, pp. 635-662.

Klute, A., Dirksen, C., 1986. Hydraulic Conductivity and Diffusivity: Laboratory Methods, In: Klute, A. (Ed.), Methods of Soil Analysis: Part 1—Physical and Mineralogical Methods. Soil Science Society of America, American Society of Agronomy, Madison, WI, pp. 687-734.

Koekkoek, E., Booltink, H., 1999. Neural network models to predict soil water retention. Eur. J. Soil Sci. 50, 489-495.

Lamorski, K., Pachepsky, Y., Sławiński, C., Walczak, R., 2008. Using support vector machines to develop pedotransfer functions for water retention of soils in Poland. Soil Sci. Soc. Am. J. 72, 1243-1247.

Lamorski, K., Sławiński, C., Moreno, F., Barna, G., Skierucha, W., Arrue, J.L., 2014. Modelling soil water retention using support vector machines with genetic algorithm optimisation. Sci. World J. 2014, 740521, 1-10.

Liaw, A., Wiener, M., 2002. Classification and regression by random forest. R news 2, 18-22.

Ließ, M., Glaser, B., Huwe, B., 2012. Uncertainty in the spatial prediction of soil texture: comparison of regression tree and Random Forest models. Geoderma 170, 70-79.

Liu, Y., 2014. Random forest algorithm in big data environment. Comput. Model. New Tech. 18, 147-151.

Ma, Y., Cukic, B., Singh, H., 2005. A classification approach to multi-biometric score fusion, International Conference on Audio-and Video-Based Biometric Person Authentication. Springer, pp. 484-493.

MathWorks, 2018. MATLAB: the language of technical computing, Inc., Natick, Massachusetts, United States.

Matin, S., Chelgani, S.C., 2016. Estimation of coal gross calorific value based on various analyses by random forest method. Fuel 177, 274-278.

Medrado, E., Lima, J.E., 2014. Development of pedotransfer functions for estimating water retention curve for tropical soils of the Brazilian savanna. Geoderma Regional 1, 59-66.

Merdun, H., Çınar, Ö., Meral, R., Apan, M., 2006. Comparison of artificial neural network and regression pedotransfer functions for prediction of soil water retention and saturated hydraulic conductivity. Soil Tillage Res. 90, 108-116.

Minasny, B., McBratney, A.B., Bristow, K.L., 1999. Comparison of different approaches to the development of pedotransfer functions for water-retention curves. Geoderma 93, 225-253.

Mualem, Y., 1976. A new model for predicting the hydraulic conductivity of unsaturated porous media. Water Resour. Res. 12, 513-522.

Nemes, A., Rawls, W.J., Pachepsky, Y.A., 2005. Influence of organic matter on the estimation of saturated hydraulic conductivity. Soil Sci. Soc. Am. J. 69, 1330-1337.

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

Nemes, A., Schaap, M., Wösten, J., 2003. Functional evaluation of pedotransfer functions derived from different scales of data collection. Soil Sci. Soc. Am. J. 67, 1093-1102.

Neyshaburi, M.R., Bayat, H., Mohammadi, K., Nariman-Zadeh, N., Irannejad, M., 2015. Improvement in estimation of soil water retention using fractal parameters and multiobjective group method of data handling. Arch. Agron. Soil Sci. 61, 257-273.

Nguyen, P.M., Haghverdi, A., De Pue, J., Botula, Y.-D., Le, K.V., Waegeman, W., Cornelis, W.M., 2017. Comparison of statistical regression and data-mining techniques in estimating soil water retention of tropical delta soils. Biosyst. Eng. 153, 12-27.

Pachepsky, Y., Rawls, W., Gimenez, D., Watt, J., 1998. Use of soil penetration resistance and group method of data handling to improve soil water retention estimates. Soil Tillage Res. 49, 117-126.

Pachepsky, Y.A., Rawls, W., 1999. Accuracy and reliability of pedotransfer functions as affected by grouping soils. Soil Sci. Soc. Am. J. 63, 1748-1757.

Pachepsky, Y.A., Rawls, W., Lin, H., 2006. Hydropedology and pedotransfer functions. Geoderma 131, 308-316.

Pachepsky, Y.A., Timlin, D., Varallyay, G., 1996. Artificial neural networks to estimate soil water retention from easily measurable data. Soil Sci. Soc. Am. J. 60, 727-733.

Rab, M., Chandra, S., Fisher, P., Robinson, N., Kitching, M., Aumann, C., Imhof, M., 2011. Modelling and prediction of soil water contents at field capacity and permanent wilting point of dryland cropping soils. Soil Res. 49, 389-407.

Rajkai, K., Kabos, S., Van Genuchten, M.T., 2004. Estimating the water retention curve from soil properties: comparison of linear, nonlinear and concomitant variable methods. Soil Tillage Res. 79, 145-152.

Rawls, W., Brakensiek, D., Soni, B., 1983. Agricultural management effects on soil water processes part I: Soil water retention and Green and Ampt infiltration parameters. Transactions of the ASAE 26, 1747-1752.

Rawls, W., Gish, T., Brakensiek, D., 1991. Estimating soil water retention from soil physical properties and characteristics, Advances in soil science. Springer, pp. 213-234.

Rawls, W.J., Brakensiek, D., 1985. Prediction of soil water properties for hydrologic modeling, Watershed management in the eighties. ASCE, pp. 293-299.

Schaap, M.G., Leij, F.J., van Genuchten, M.T., 1998. Neural network analysis for hierarchical prediction of soil hydraulic properties. Soil Sci. Soc. Am. J. 62, 847-855.

Schaap, M.G., Leij, F.J., van Genuchten, M.T., 2001. Rosetta: A computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. J. Hydrol. 251, 163-176.

Scheinost, A., Sinowski, W., Auerswald, K., 1997. Regionalization of soil water retention curves in a highly variable soilscape, I. Developing a new pedotransfer function. Geoderma 78, 129-143.

Seo, S., 2006. A review and comparison of methods for detecting outliers in univariate data sets, Thesis for Master of Science in Field of Public Health University of Pittsburgh, pp. 1-59.

Shirazi, M.A., Boersma, L., 1984. A unifying quantitative analysis of soil texture. Soil Sci. Soc. Am. J. 48, 142-147.

Shwetha, P., Varija, K., 2015. Soil water retention curve from saturated hydraulic conductivity for sandy loam and loamy sand textured soils. Aquat. Procedia 4, 1142-1149.

Simmons, J.P., Nelson, L.D., Simonsohn, U., 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol. Sci. 22, 1359-1366.

Szabó, B., Szatmári, G., Takács, K., Laborczi, A., Makó, A., Rajkai, K., Pásztor, L., 2019. Mapping soil hydraulic properties using random forest based pedotransfer functions and geostatistics. Hydrol. Earth Syst. Sci. 23, 2615-2635.

Tietje, O., Tapkenhinrichs, M., 1993. Evaluation of pedo-transfer functions. Soil Sci. Soc. Am. J. 57, 1088-1095.

Tomasella, J., Hodnett, M.G., 1998. Estimating soil water retention characteristics from limited data in Brazilian Amazonia. Soil Sci. 163, 190-202.

Tomasella, J., Hodnett, M.G., Rossato, L., 2000. Pedotransfer functions for the estimation of soil water retention in Brazilian soils. Soil Sci. Soc. Am. J. 64, 327-338.

Tóth, B., Makó, A., Toth, G., 2014. Role of soil properties in water retention characteristics of main Hungarian soil types. J. Cent. Eur. Agric. 15, 137-153.

Touil, S., Degre, A., Chabaca, M.N., 2016. Sensitivity analysis of point and parametric pedotransfer functions for estimating water retention of soils in Algeria. Soil 2, 647.

Twarakavi, N.K., Šimůnek, J., Schaap, M., 2009. Development of pedotransfer functions for estimation of soil hydraulic parameters using support vector machines. Soil Sci. Soc. Am. J. 73, 1443-1452.

Ungaro, F., Calzolari, C., Busoni, E., 2005. Development of pedotransfer functions using a group method of data handling for the soil of the Pianura Padano–Veneta region of North Italy: water retention properties. Geoderma 124, 293-317.

van Genuchten, M.T., 1980. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. Soil Sci. Soc. Am. J. 44, 892-898.

Verhagen, J., 1997. Site specific fertiliser application for potato production and effects on N-leaching using dynamic simulation modelling. Agric. Ecosyst. Environ. 66, 165-175.

Verikas, A., Gelzinis, A., Bacauskiene, M., 2011. Mining data with random forests: A survey and results of new tests. Pattern Recognit. 44, 330-349.

Walkley, A., Black, I.A., 1934. An examination of the Degtjareff method for determining soil organic matter, and a proposed modification of the chromic acid titration method. Soil Sci. 37, 29-38.

Wassar, F., Gandolfi, C., Rienzner, M., Chiaradia, E.A., Bernardoni, E., 2016. Predicted and measured soil retention curve parameters in Lombardy region north of Italy. International Soil and Water Conservation Research 4, 207-214.

Wiesmeier, M., Barthold, F., Blank, B., Kögel-Knabner, I., 2011. Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. Plant Soil 340, 7-24.

Williams, J., Ross, P., Bristow, K.L., 1992. Prediction of the Campbell water retention function from texture, structure, and organic matter. In 'Indirect methods for estimating the hydraulic properties of unsaturated soils.' University of California: Riverside.

Wösten, J., Pachepsky, Y.A., Rawls, W., 2001. Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic characteristics. J. Hydrol. 251, 123-150.

Zaklouta, F., Stanciulescu, B., 2012. Real-time traffic-sign recognition using tree classifiers. IEEE Transactions on Intelligent Transportation Systems 13, 1507-1514.

Zhao, P., Su, X., Ge, T., Fan, J., 2016. Propensity score and proximity matching using random forest. Contemp. Clin. Trials 47, 85-92.

Figure captions

Fig 1. Input variables of the 15 pedotransfer functions (PTFs) for predicting the van Genuchten model parameters (θ_r, θ_s, α and n) of the soil water retention curve (SWRC). A list of abbreviations is available in the notation box.

Fig. 2. An architecture of a random forest.

Fig. 3. Variation of soil texture classes for the dataset (n = 223) on the United States Department of Agriculture (USDA) textural triangle.

Fig. 4. Correlation matrix plot between input and output variables.

** Correlation is significant at the P<0.01 level.

* Correlation is significant at the P<0.05 level.

A list of abbreviations is available in the notation box.

Fig. 5. Results of the prediction of the soil water retention curve (SWRC) through the van Genuchten model by the nonlinear regression (NLR) and random forests (RF) techniques for the training step as reflected in the integral mean error (*IME*), integral root mean square error (*IRMSE*), coefficient of determination (R_2), and Akaike's information criterion (*AIC*). Vertical lines indicate the standard deviations. Means with the same letter are not significantly different at the significance level of P<0.05 (*IRMSE* only).

Fig. 6. Results of the prediction of the soil water retention curve (SWRC) through the van Genuchten model by the Rosetta software, nonlinear regression (NLR) and random forests (RF) techniques for the testing step as reflected in the integral mean error (*IME*), integral root mean square error (*IRMSE*), coefficient of determination (R_2), and Akaike's information criterion (*AIC*). Vertical lines indicate the standard deviations. Means with the same letter are not significantly different at the significance level of P<0.05 (*IRMSE* only).

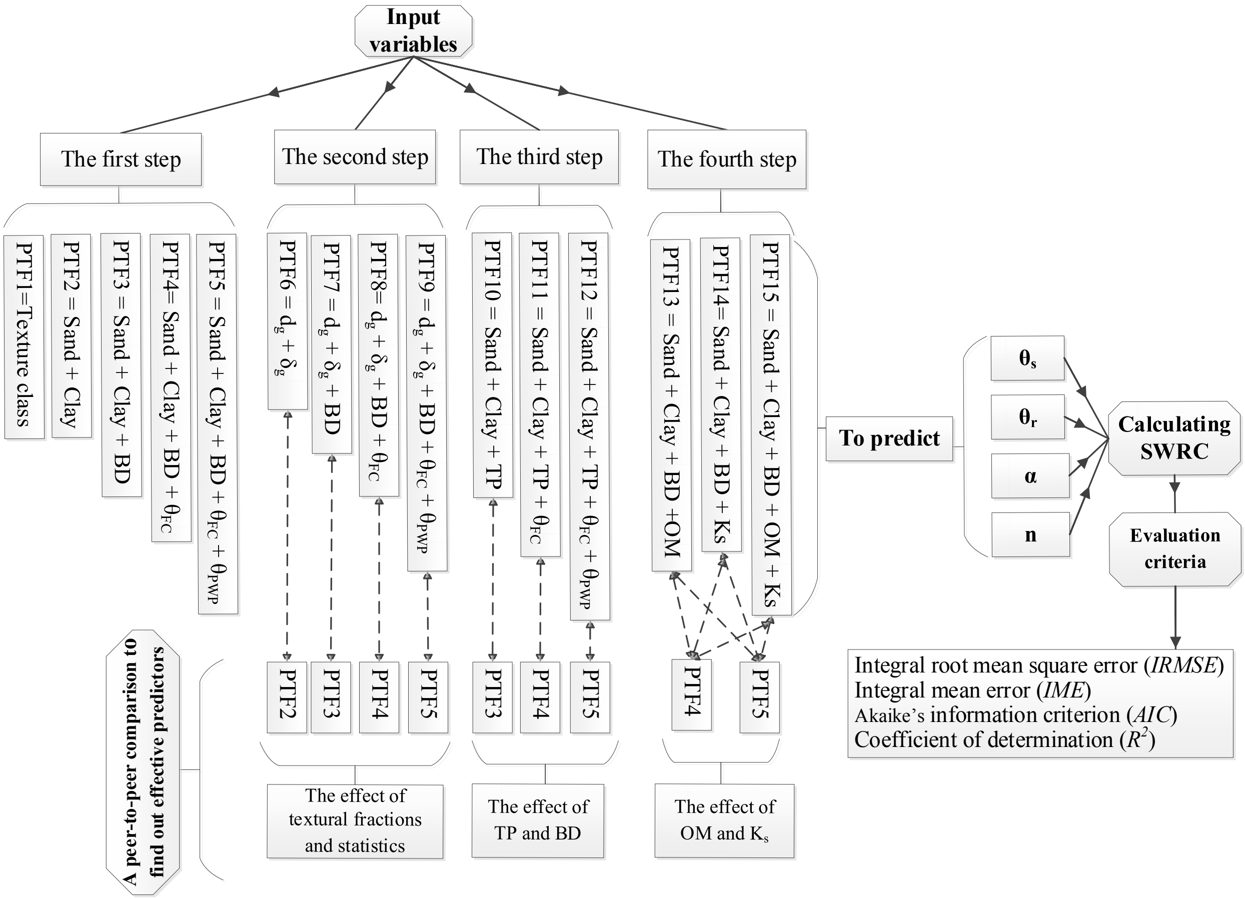

Input variables
The first step
The second step
The third step
The fourth step
PTF1=Texture class
PTF2 = Sand + Clay
PTF3 = Sand + Clay + BD
PTF4= Sand + Clay + BD + θ_{FC}
PTF5 = Sand + Clay + BD + θ_{FC} + θ_{PWP}
PTF6 = d_g + δ_g
PTF7 = d_g + δ_g + BD
PTF8= d_g + δ_g + BD + θ_{FC}
PTF9 = d_g + δ_g + BD + θ_{FC} + θ_{PWP}
PTF10 = Sand + Clay + TP
PTF11 = Sand + Clay + TP + θ_{FC}
PTF12 = Sand + Clay + TP + θ_{FC} + θ_{PWP}
PTF13 = Sand + Clay + BD +OM
PTF14= Sand + Clay + BD + Ks
PTF15 = Sand + Clay + BD + OM + Ks
To predict
θ_s
θ_r
α
n
Calculating SWRC
Evaluation criteria
Integral root mean square error (*IRMSE*)
Integral mean error (*IME*)
Akaike's information criterion (*AIC*)
Coefficient of determination (R^2)
PTF2
PTF3
PTF4
PTF5
PTF3
PTF4
PTF5
PTF4
PTF5
A peer-to-peer comparison to find out effective predictors
The effect of textural fractions and statistics
The effect of TP and BD
The effect of OM and K_s

Fig 1. Input variables of the 15 pedotransfer functions (PTFs) for predicting the van Genuchten model parameters (θ_r, θ_s, α and n) of the soil water retention curve (SWRC). A list of abbreviations is available in the notation box.

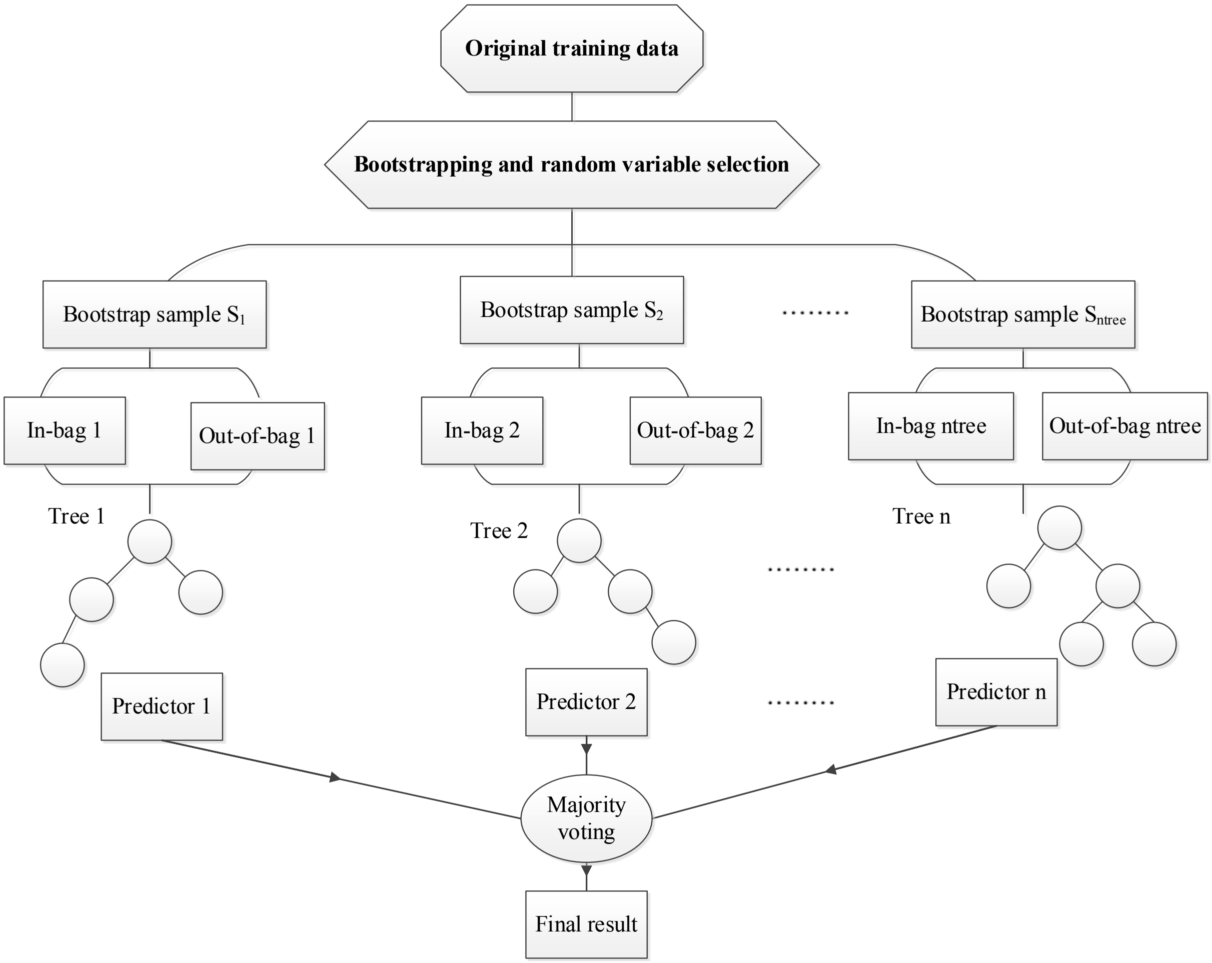


Fig. 2. An architecture of a random forest.

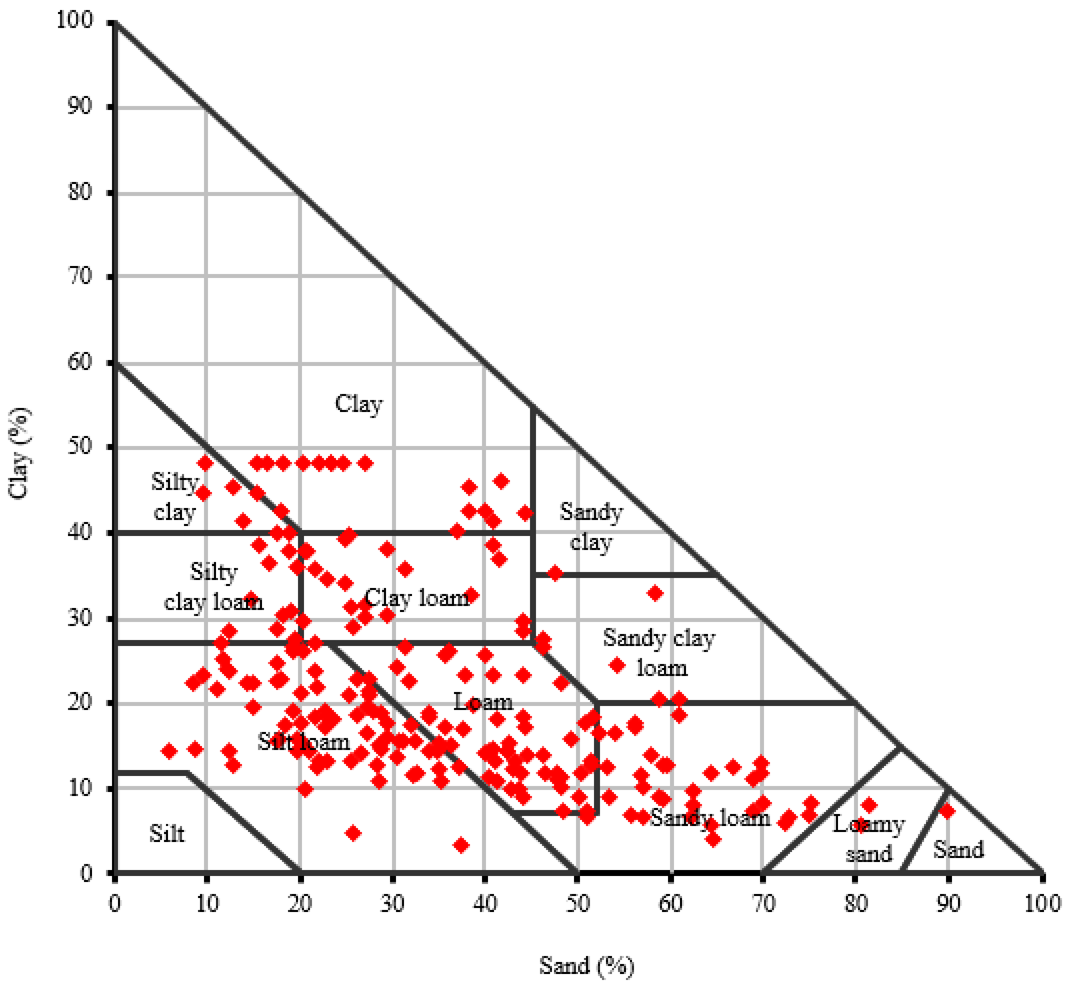


Fig. 3. Variation of soil texture classes for the dataset (n = 223) on the United States Department of Agriculture (USDA) textural triangle.

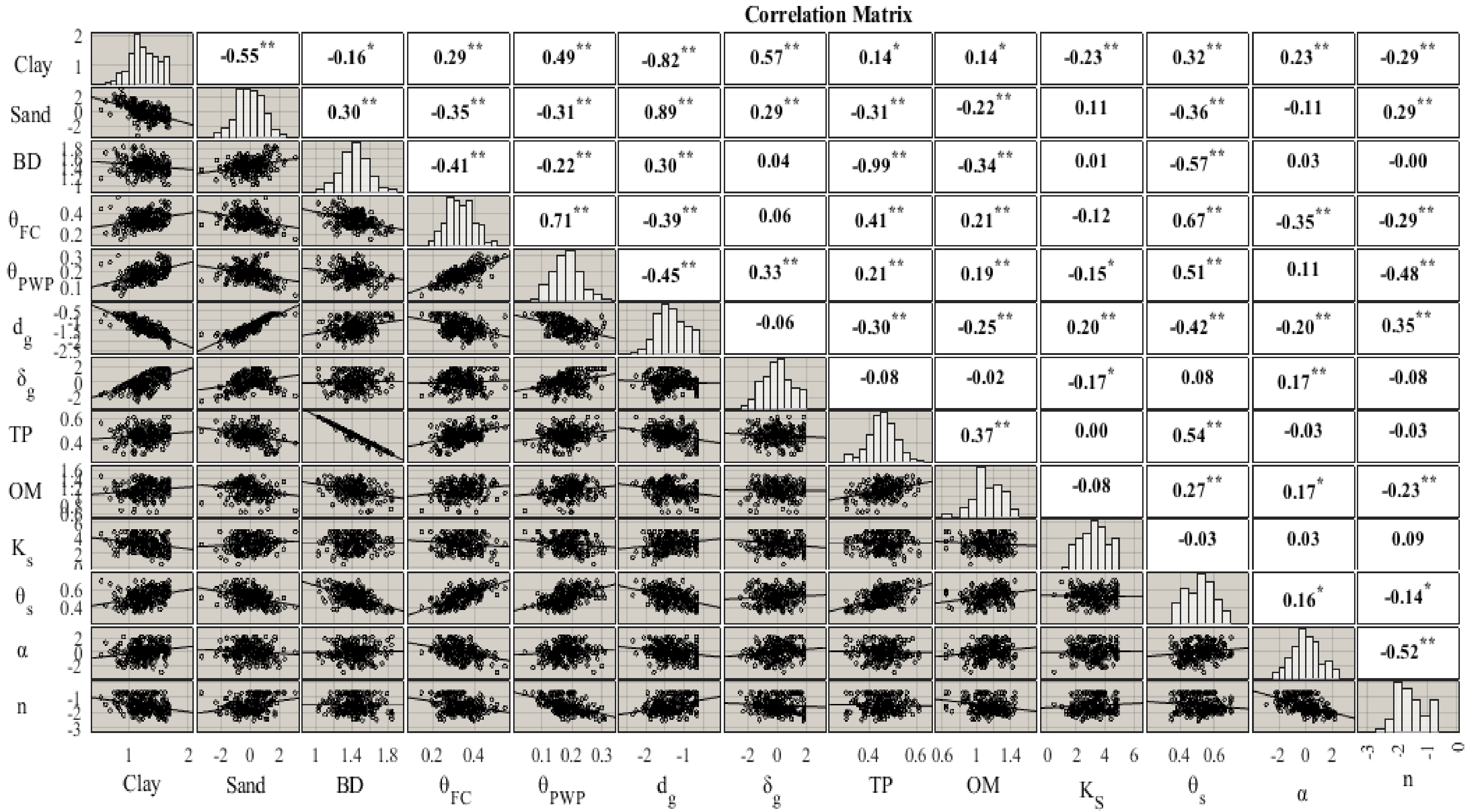
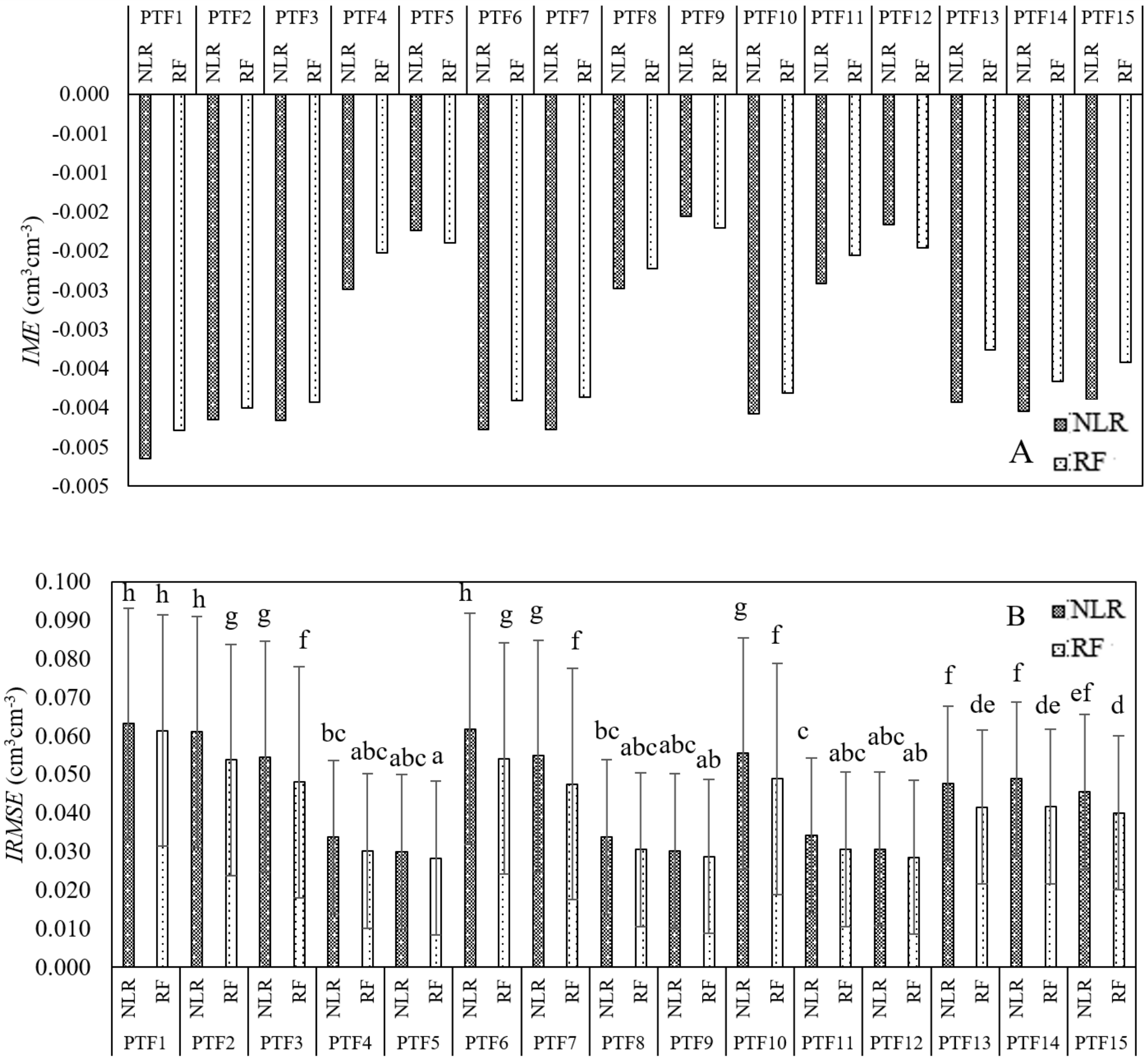


Fig. 4. Correlation matrix plot between input and output variables.

** Correlation is significant at the *P*<0.01 level.

* Correlation is significant at the *P*<0.05 level.

A list of abbreviations is available in the notation box.


PTF1
PTF2
PTF3
PTF4
PTF5
PTF6
PTF7
PTF8
PTF9
PTF10
PTF11
PTF12
PTF13
PTF14
PTF15
NLR
RF
0.000
-0.001
-0.001
-0.002
-0.002
-0.003
-0.003
-0.004
-0.004
-0.005
-0.005
IME (cm³cm⁻³)
NLR
RF
A
0.100
0.090
0.080
0.070
0.060
0.050
0.040
0.030
0.020
0.010
0.000
IRMSE (cm³cm⁻³)
B
NLR
RF
h
h
h
g
g
f
bc
abc
abc
a
h
g
g
f
bc
abc
abc
ab
g
f
c
abc
abc
ab
f
de
f
de
ef
d

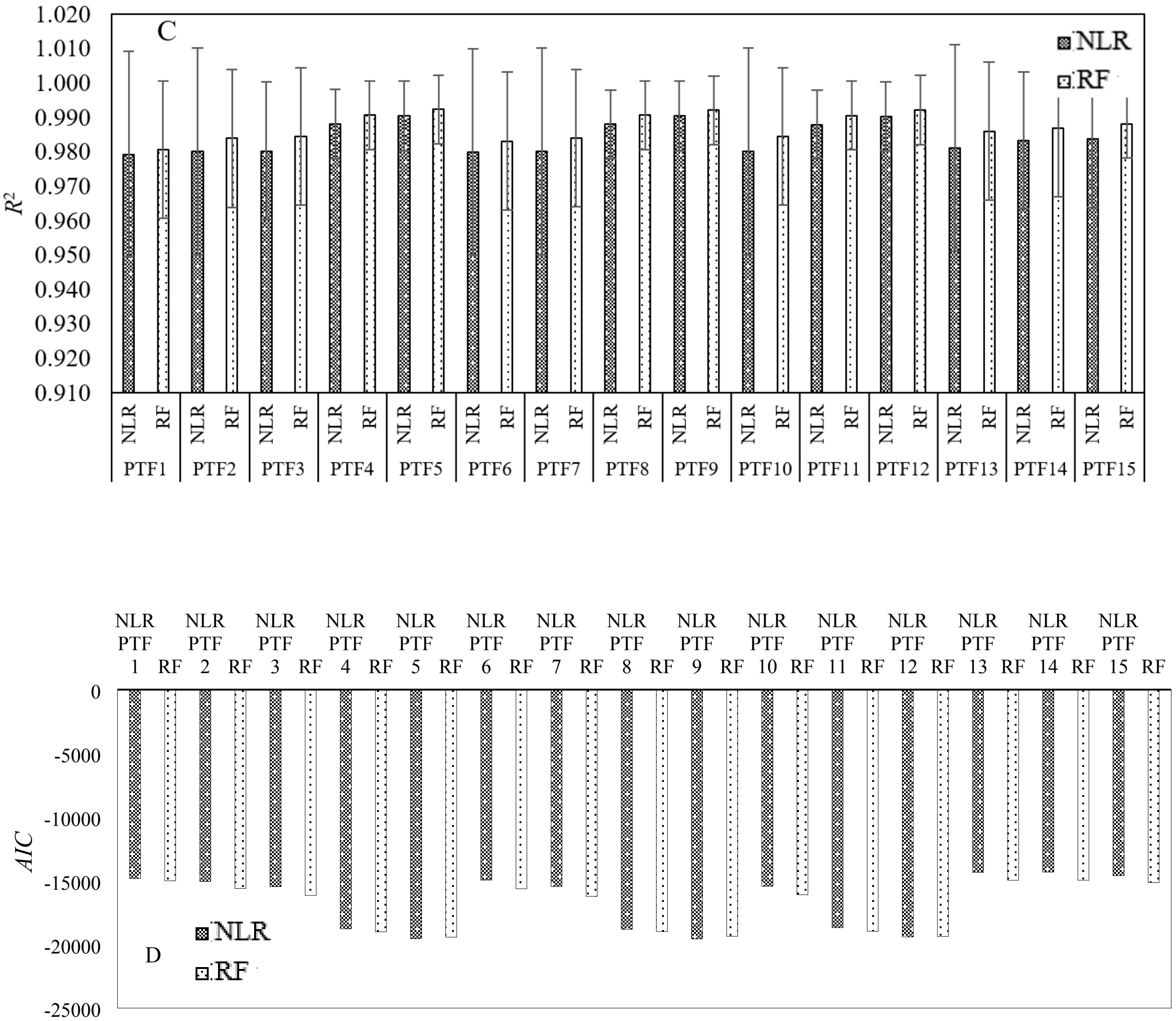
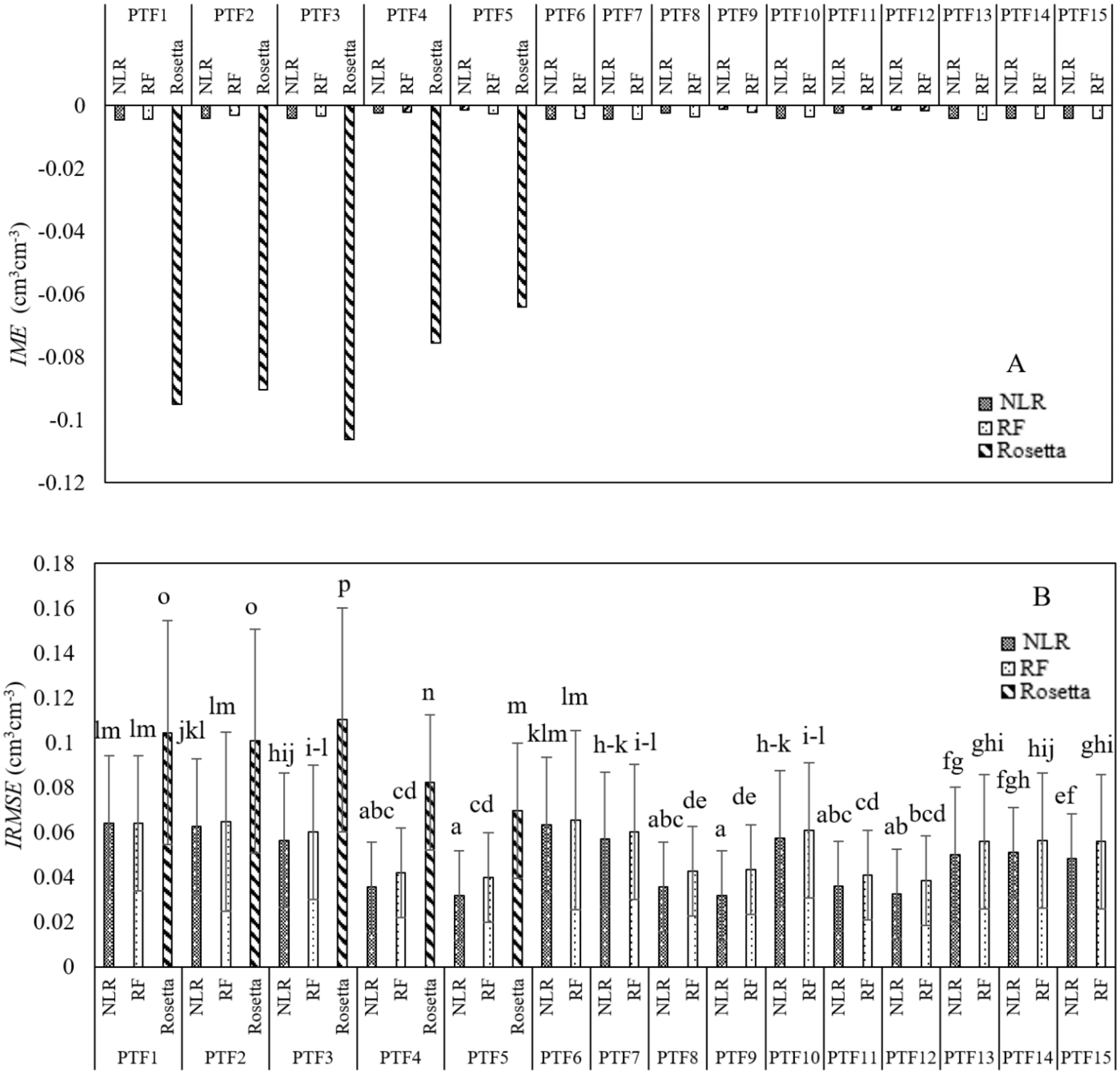


Fig. 5. Results of the prediction of the soil water retention curve (SWRC) through the van Genuchten model by the nonlinear regression (NLR) and random forests (RF) techniques for the training step as reflected in the integral mean error (*IME*), integral root mean square error (*IRMSE*), coefficient of determination (R_2), and Akaike's information criterion (*AIC*). Vertical lines indicate the standard deviations. Means with the same letter are not significantly different at the significance level of $P<0.05$ (*IRMSE* only).


PTF1
PTF2
PTF3
PTF4
PTF5
PTF6
PTF7
PTF8
PTF9
PTF10
PTF11
PTF12
PTF13
PTF14
PTF15
NLR
RF
Rosetta
0
-0.02
-0.04
-0.06
-0.08
-0.1
-0.12
IME (cm^{3}cm^{-3})
A
▨ NLR
□ RF
◪ Rosetta
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
$IRMSE$ (cm^{3}cm^{-3})
B
▨ NLR
□ RF
◪ Rosetta
lm
lm
o
jkl
lm
o
hij
i-l
p
abc
cd
n
a
cd
m
klm
lm
h-k
i-l
abc
de
a
de
h-k
i-l
abc
cd
ab
bcd
fg
ghi
fgh
hij
ef
ghi
NLR
RF
Rosetta
PTF1
PTF2
PTF3
PTF4
PTF5
PTF6
PTF7
PTF8
PTF9
PTF10
PTF11
PTF12
PTF13
PTF14
PTF15

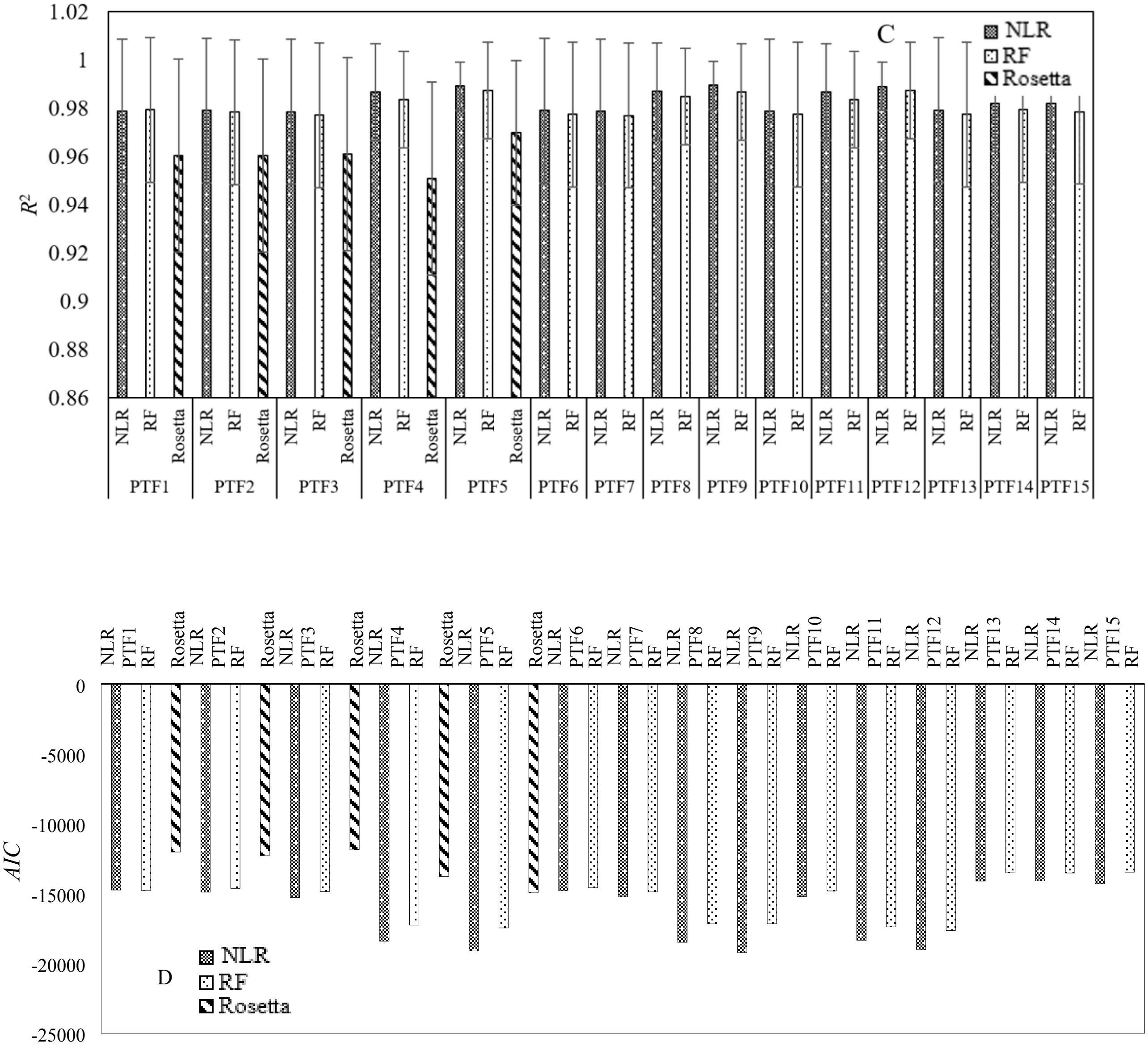


Fig. 6. Results of the prediction of the soil water retention curve (SWRC) through the van Genuchten model by the Rosetta software, nonlinear regression (NLR) and random forests (RF) techniques for the testing step as reflected in the integral mean error (*IME*), integral root mean square error (*IRMSE*), coefficient of determination (R_2), and Akaike's information criterion (*AIC*). Vertical lines indicate the standard deviations. Means with the same letter are not significantly different at the significance level of *P*<0.05 (*IRMSE* only).

Table 1- The results of 10, 15 and 20-fold cross-validation (k) for van Genuchten model parameters of the soil water retention curve derived from nonlinear regression (NLR) and random forest (RF) techniques based on root mean square error (*RMSE*) for pedotransfer functions PTF 3, 5 and 11 in the train and test datasets.

			θ_r			θ_s			α			n		
			RMSE			*RMSE*			*RMSE*			*RMSE*		
			Train	Test	Mean	Train	Test	Mean	Train	Test	Mean	Train	Test	Mean
PTF3	k=10	NLR	0.058	0.060	0.059	0.063	0.065	0.064	1.017	1.037	1.027	0.426	0.436	0.431
		RF	0.052	0.061	0.056	0.058	0.073	0.066	0.893	1.084	0.989	0.374	0.442	0.408
	k=15	NLR	0.058	0.060	0.059	0.064	0.064	0.064	1.017	1.030	1.024	0.426	0.434	0.430
		RF	0.052	0.061	0.057	0.058	0.070	0.064	0.894	1.033	0.964	0.374	0.441	0.408
	k=20	NLR	0.058	0.060	0.059	0.064	0.064	0.064	0.064	0.064	0.064	0.426	0.437	0.432
		RF	0.051	0.060	0.056	0.057	0.071	0.064	0.057	0.071	0.064	0.368	0.442	0.405
PTF5	k=10	NLR	0.051	0.053	0.052	0.053	0.054	0.054	0.764	0.796	0.780	0.380	0.397	0.389
		RF	0.043	0.056	0.050	0.046	0.056	0.051	0.675	0.869	0.772	0.327	0.411	0.369
	k=15	NLR	0.051	0.053	0.052	0.053	0.055	0.054	0.764	0.790	0.777	0.381	0.399	0.390
		RF	0.044	0.054	0.049	0.046	0.055	0.050	0.679	0.848	0.763	0.329	0.421	0.375
	k=20	NLR	0.051	0.053	0.052	0.053	0.055	0.054	0.765	0.789	0.777	0.381	0.399	0.390
		RF	0.042	0.054	0.048	0.044	0.054	0.049	0.654	0.842	0.748	0.316	0.412	0.364
PTF11	k=10	NLR	0.058	0.061	0.060	0.065	0.067	0.066	1.018	1.052	1.035	0.431	0.448	0.440
		RF	0.050	0.061	0.056	0.047	0.057	0.052	0.770	0.978	0.874	0.370	0.443	0.406
	k=15	NLR	0.058	0.061	0.060	0.065	0.067	0.066	1.019	1.037	1.028	0.432	0.447	0.439
		RF	0.050	0.060	0.055	0.047	0.057	0.052	0.770	1.009	0.889	0.369	0.450	0.410
	k=20	NLR	0.058	0.060	0.059	0.065	0.065	0.065	1.020	1.024	1.022	0.432	0.439	0.435
		RF	0.049	0.061	0.055	0.046	0.056	0.051	0.745	0.964	0.855	0.361	0.443	0.402

Table 2- Some descriptive statistics of the measured soil variables and parameters of the van Genuchten model of the soil water retention curve for the entire dataset (223 soil samples).

Variables[a]	Mean	CV (%)	Minimum	Maximum	*P*-value
Clay content (%)	21.39	54.05	3.47	48.00	0.00
Log (clay content)	1.27	19.08	0.54	1.68	0.66
Sand content (%)	35.45	48.40	5.90	89.80	0.00
Sand content*	-0.01	-14350.94	-3.40	3.14	0.90
Bulk density (g cm^{-3})	1.43	10.97	1.03	1.84	0.83
θ_{FC} (cm^3 cm^{-3})[$]	0.33	20.44	0.15	0.55	0.45
θ_{PWP} (cm^3 cm^{-3})	0.18	26.21	0.04	0.31	0.90
d_g (mm)	0.07	86.62	0.00	0.21	0.00
Log (d_g)	-1.33	-27.91	-2.34	-0.67	0.77
δ_g (-)	11.57	29.39	4.54	19.97	0.00
δ_g*	-0.01	-9872.87	-2.53	1.80	0.96
Total porosity (cm^3 cm^{-3})	0.46	13.26	0.31	0.61	0.67
Organic matter content (%)	1.84	53.68	0.17	4.41	0.00
(Organic matter content)$^{(1/4)}$	1.13	14.83	0.64	1.45	0.86
K_s (cm day^{-1})	169.10	96.58	0.06	530	0.00
$(K_s)^{(1/4)}$	3.23	30.37	0.50	4.80	0.59
θ_r (cm^3 cm^{-3})	0.04	158.05	0.00	0.17	0.00
θ_s (cm^3 cm^{-3})	0.52	16.26	0.35	0.70	0.56
α (kPa^{-1})	0.06	115.62	0.00	0.29	0.00
α*	0.01	8889.14	-2.93	2.19	0.93
n	1.24	9.80	1.08	1.48	0.00
Ln (n-1)	-1.55	-30.92	-2.52	-0.74	0.05

[a] CV, coefficient of variation.

[$]. A list of abbreviations is available in the notation box.

* Normalized form of sand content: 0.91+1.06×Ln((sand content- 4.3)/(100.2-sand content)); normalized form of δ_g: -1.04657+1.39359×Asinh((δ_g- 8.4)/3.04); and normalized form of α: 3.6+0.92×Ln((α- 8.2×10^{-6})/(1.6-α)). *P*-value is a significance value for normality test.

Table 3- The variance inflation factor (*VIF*) values for normalized form of the input variables.

PTFs	Clay* (%)	Sand (%)	BD$ (g cm^{-3})	θ_{FC} (cm^3 cm^{-3})	θ_{PWP} (cm^3 cm^{-3})	d_g (mm)	δ_g (-)	TP (cm^3 cm^{-3})	OM (%)	K_s (cm day^{-1})
PTF2	1.42	1.42								
PTF3	1.43	1.52	1.10							
PTF4	1.45	1.56	1.25	1.31						
PTF5	1.79	1.58	1.27	2.48	2.56					
PTF6						1.00	1.00			
PTF7			1.11			1.11	1.01			
PTF8			1.25	1.33		1.01	1.22			
PTF9			1.28	2.50	2.73	1.34	1.22			
PTF10	1.55	1.43						1.11		
PTF11	1.58	1.46		1.32				1.26		
PTF12	1.60	1.79		2.49	2.56			1.28		
PTF13	1.48	1.65	1.25						1.14	
PTF14	1.55	1.64	1.14							1.06
PTF15	1.55	1.65	1.25						1.15	1.06

* Normalized form of the input variables is available in Table 2.

$. A list of abbreviations is available in the notation box.

Table 4- Analysis of variance of the integral root mean square error (*IRMSE*) of the prediction of the soil water retention curve by different methods (nonlinear regression and random forest) and pedotransfer functions (PTFs 1-15) for both the train and test datasets.

	Source	Degree freedom	Mean square	*F*-value	*P*-value
Train	Repeat (Block)	222	0.007	19.09	<0.0001
	PTFs	14	0.062	180.68	<0.0001
	Methods	1	0.038	109.69	<0.0001
	PTFs × Methods	14	0.001	1.78	0.0356
	Error	6288	0.0003		
Test	Repeat (Block)	222	0.010	16.04	<0.0001
	PTFs	14	0.073	117.22	<0.0001
	Methods	2	0.656	1056.43	<0.0001
	PTFs × Methods	18	0.002	3.68	<0.0001
	Error	7398	0.0006		

Declaration of interests

▯ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Author statement:

Mostafa Rastgou:

Data curation, Writing- Original draft preparation, Visualization, Investigation, Formal analysis.

Hossein Bayat:

Conceptualization, Methodology, Writing, Supervision, Project administration, Funding acquisition.

Muharram Mansoorizadeh:

Software, Validation.

Andrew S. Gregory:

Writing- Reviewing and Editing