Geoderma 226–227 (2014) 31–38

Contents lists available at ScienceDirect

Geoderma

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

Revisiting the pseudo continuous pedotransfer function concept: Impact of data quality and data mining method

A. Haghverdi [a,*], H.S. Öztürk [b], W.M. Cornelis [c]

[a] Dept. of Biosystems Engineering and Soil Science, University of Tennessee, 2506 E.J. Chapman Dr., Knoxville, TN 37996-4531, USA
[b] Department of Soil Science, Faculty of Agriculture, Ankara University, 06110 Diskapi, Ankara, Turkey
[c] Department of Soil Management, Ghent University, Coupure links 653, B-9000 Ghent, Belgium

ARTICLE INFO

Article history:
Received 10 June 2013
Received in revised form 18 February 2014
Accepted 28 February 2014
Available online 21 March 2014

Keywords:
Neural network
Pseudo continuous PTF
Support vector machine
Water retention curve

ABSTRACT

The pedotransfer function (PTF) concept has been widely used in recent years as an indirect way to predict soil hydraulic properties, particularly the water retention curve (WRC). The pseudo continuous (PC) approach allows us to predict water content at any predefined matric head, resulting in an almost continuous WRC. When combined with powerful pattern recognition approaches, a PC-PTF can be trained to learn the shape of WRC from a discrete set of measured points, unlike traditional parametric PTFs which follow a predefined WRC shape dictated by the selected soil hydraulic equations. The purpose of this study was to investigate the impact of two elements on the performance of a PC-PTF: (i) data mining method (neural network, NN, versus support vector machine, SVM) and (ii) distribution and density of the provided water retention data in the training phase. Two datasets from Turkey and Belgium, consisting of mainly fine and coarse-textured soils, respectively, were employed. Multiple scenarios containing different combinations of measured water retention points in the training phase were defined. The lower root mean square error (RMSE) on average (0.044 cm^3 cm^{-3}) attained with the NN-based PTF shows that it is a better option than SVM (RMSE of 0.052 cm^3 cm^{-3}) for deriving PC-PTFs. The accuracy of PC-PTF was firmly dependent on the presence of measured water retention points in the entire range of WRC. Applying different scenarios revealed that a well distributed set of measured water retention points in the training phase could result in up to 0.03 cm^3 cm^{-3} reduction in RMSE values.

1. Introduction

A majority of agro-hydrological studies requires information on unsaturated soil hydraulic characteristics such as water retention and hydraulic conductivity. However, due to their significant temporal and spatial variability, these properties are classified as hardly available data. Direct methods of measurement are still time consuming and costly to be practical at large scales even though major advancement has been achieved in the development of lab and field measurement techniques (Vereecken et al., 2010). Therefore, deriving pedotransfer functions (PTFs) for making a mathematical connection between easily collected data and hydraulic properties is still a demanding research area.

Traditionally, PTFs enable to predict either individual points on the water retention curve (WRC) (i.e., a point PTF) or parameters of a soil hydraulic equation that describes WRC (i.e. a parametric PTF) (Wösten et al., 2001). Inherent handicaps of parametric PTFs are: (i) the real shape of WRC is not always identical to the one of the selected equation and (ii) the WRC formed by outputs of the parametric PTF always carries more error than the fitted WRC. In both cases, i.e., WRC fitting and PTF derivation, the soil hydraulic equation could be withdrawn and replaced by an appropriate neural network (NN) model (Haghverdi et al., 2012; Jain et al., 2004). Jain et al. (2004) proved that the performance of NN for WRC fitting is similar and in some cases even better than the performance of the van Genuchten equation, VG (van Genuchten, 1980).

NN was the first data mining (DM) method, employed by Schaap and Bouten (1996) and Pachepsky et al. (1996), for developing a PTF. Successful results of NN-PTF encouraged researchers to examine novel DM algorithms. Among the DM based techniques, support vector machine (SVM) is claimed to be more capable than NN to overcome local minimum and overfitting problems (Cristianini and Shawe-Taylor, 2000; Vapnik, 1995). SVM is a supervised learning method based on statistical learning theory. SVM is among the top 10 DM methods which were applied in different agricultural related studies (Mucherino et al., 2009). There are also some studies illustrating that SVM may be a better candidate than NN either for deriving a point (Lamorski et al., 2008) or a parametric PTF (Twarakavi et al., 2009). Lamorski et al. (2008) found that SVM had the same accuracy as NN and in some cases even outperformed NN. Hence, it is beneficial to consider SVM as a tool for

* Corresponding author. Tel.: +1 8652359694 (mobile).
E-mail addresses: ahaghver@utk.edu (A. Haghverdi), wim.cornelis@UGent.be (W.M. Cornelis).

http://dx.doi.org/10.1016/j.geoderma.2014.02.026
0016-7061/

PTF development. In another study, Twarakavi et al. (2009) compared SVM with NN (i.e. Rosetta-PTF) and obtained higher accuracy by SVM.

Recently Haghverdi et al. (2012) introduced a new form of NN-based PTF, named pseudo continuous (PC). It is capable of determining almost continuous WRCs without using any soil hydraulic equation. They demonstrated that the PC-NN-based PTF ($PC_{NN}PTF$) could be more accurate and reliable than parametric PTFs based on VG equation. The $PC_{NN}PTF$ could be utilized for predicting water contents at some specific points, such as e.g., field capacity, or for predicting a complete WRC. Despite the promising performance of the $PC_{NN}PTF$, there are still some crucial unanswered questions on how to apply this new method to predict WRC.

In theory, the $PC_{NN}PTF$ is able to predict water content at any desired matric potential, whereas Haghverdi et al. (2012) mostly used and evaluated it as a point PTF. They observed an increase in error when employing the $PC_{NN}PTF$ for matric potentials different than those available in the training dataset. One of the unique characteristic of the $PC_{NN}PTF$ is its structure that enables establishing a PTF from a dataset containing samples with uneven measured points. However, the $PC_{NN}PTF$ is very sensitive to density and distribution of the provided water retention points in the training process because it learns the shape of WRC from data. Therefore, performance of the $PC_{NN}PTF$ should be carefully investigated for continuous prediction of WRC. In addition, since PC-PTF only utilizes the power of the DM method to predict the shape of WRC, its performance seems to be closely related to the selected DM method.

To derive a new PTF, a dataset should be established with sufficient amount of information on desired soil properties. One possible option is to form a big database by combining some available datasets from different origins. For instance Botula et al. (2013) and Twarakavi et al. (2009) utilized a selected subset of IGBP-DIS international database from ISRIC (Tempel et al., 1996) and UNSODA database (Nemes et al., 2001) to derive their PTFs, respectively. Nemes (2011) reviewed available international databases of soil physical and hydraulic properties. Alternatively, time-consuming sampling and measuring campaigns can be setup to establish a smaller local database. In developing countries, soil databases are insufficient though slowly developing (Botula et al., 2013). It was already proven that using local data for deriving PTFs is a preferable option where it is desired to apply subsequently the established model for similar soils (Haghverdi et al., 2012). PC-PTF is a suitable option to derive local PTFs since it was introduced as a useful empirical model to deal with small datasets (Haghverdi et al., 2012). To tackle this goal, identifying the optimized number of required measured water retention points is extremely important since it substantially affects the associated time and cost with desired lab/field experiments as well as the accuracy and reliability of the established PTFs.

This study was carried out (i) to evaluate the performance of PC-PTF to predict a continuous WRC, (ii) to identify optimized required water retention points to derive the PC-PTF and (iii) to investigate the impact of data mining algorithm (NN and SVM) on the performance of PC-PTF.

2. Material and methods

2.1. Experimental data

Two datasets from Turkey and Belgium were employed in this study. The first dataset contained 135 disturbed and undisturbed (100 cm^3) samples from the surface soil (0–30 cm) mostly in the area surrounding Ankara, Turkey. Two undisturbed samples were taken from each location using a dedicated soil sample ring kit. The water content of samples was close to but lower than their field capacity at the time of sampling. To eliminate the effect of management practices on soil structure and in turn on water content, the sampling locations were selected from long term fallow lands and non-agricultural lands. Particle-size distribution and organic matter content (OM) were determined using disturbed soil samples by the hydrometer method (Gee and Bauder, 1986) and the modified method of Walkley and Black (Jackson, 1958), respectively. The soil water contents at 8 different matric potentials (i.e. −5, −10, −33, −100, −400, −700, −1000, −1500 kPa) were measured with sand box apparatus (Eijkelkamp Agrisearch Equipment, Giesbeek, The Netherlands) and pressure plate equipment (Soilmoisture Equipment, Santa Barbara CA, USA). The soil cores later were used to determine bulk density (BD) (Blake and Hartge, 1986).

The second dataset contained 69 soil samples including a wide range of soil textures from Belgium. Details about this dataset, instrumentation and measurements, could be found in Cornelis et al. (2001). Samples had 8 to 10 water retention data points ranging from 0 to −1500 kPa. The soil texture and physical properties for both datasets are presented in Fig. 1 and Table 1, respectively.

2.2. Pseudo continuous-PTF

The structure of the PC-PTF differs from the point and the parametric PTFs in that in addition to basic soil data, matric potential is considered as an input variable. This topology enables the user to predict water content at any desired matric potential. In the structure of the PC-PTF, water content is the only desired output variable that corresponds to the predefined input variable, matric potential. The PC-PTF can then be applied to a wide range of input matric potentials to predict the WRC.

In this study two different data mining methods, NN and SVM, were utilized to drive the PC-PTF ($PC_{NN}PTF$ and $PC_{SVM}PTF$, respectively). The routine input predictors were clay, sand, silt, BD and OM content, where the natural logarithm of the matric potential was considered as an extra input variable.

2.3. Neural network NN

A standard three layer perceptron was adopted to derive the $PC_{NN}PTF$. Tangent hyperbolic and linear were the activation functions in the hidden layer and output layer, respectively. The Levenberg–Marquardt algorithm (Demuth and Beale, 2000) was used for the network training process. All NN modeling steps were performed using the evaluation version of Neurosolution 6.0 (www.nd.com) software. To avoid overtraining, the development dataset was divided into two individual parts where 65% of the whole data was used for training and 15% for cross-validation. The remaining 20% was allocated to test the performance of the PTF. The cross-validation part stopped the training process using a supervised learning control. The maximum number of iterations over the training set was set to 1000. The training process stopped when the mean square error of the cross-validation set began to increase. The number of neurons in the hidden layer changed from 1 to 15, and the training was repeated three times for each number of the hidden neurons.

2.4. Support vector machine SVM

SVM is a machine learning method which uses a supervised learning procedure to project the input–output relationship by mapping inputs to a high-dimensional feature space, and subsequently putting them in different classes. Detailed information about the SVM theory can be found in Twarakavi et al. (2009). Similar to the majority of DM methods, SVM was first introduced to address classification problems and later to solve regression questions. The most noticeable advantage of SVM is the elimination of the local minimum issue. SVMs automatically choose their model size based on the principle of structural risk optimization. Regularization parameter, C, regression precision, ε, and γ are the parameters that ought to be optimized for achieving the best performance of SVMs. Parameter C governs the balance between model simplicity and accuracy. Parameter ε works as a threshold for the error in a manner that just the error greater than ε would be taken into account. Transformation of the input data is often carried out utilizing nonlinear

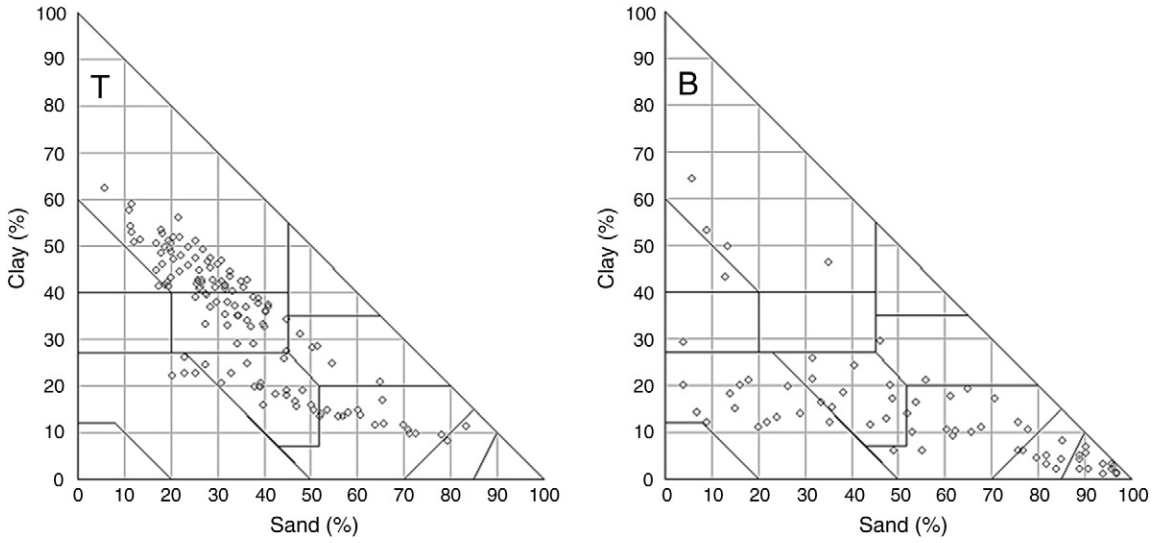

Fig. 1. Soil texture of the datasets from Turkey (left panel) and Belgium (right panel). Clay corresponds to 0–2 μm, silt to 2–50 μm and sand to 50–2000 μm.

kernel functions (Lamorski et al., 2008). Parameter γ is the kernel width of the applied kernel function.

SVM-PTF computations in this study were performed utilizing SPSS Clementine 12.0 (www.ibm.com). The radial basis function was the kernel function. Parameter C was changed between 5 and 200, and ε was varied from 0.05 to 0.5 while the range of variation for γ was from 1 to 7. The best combination of these parameters was identified utilizing a cross-validation technique. The designated data to train, cross-validation and test as well as the input and output variables were identical to the ones applied to NN.

2.5. Distribution and density of the retention points

To identify the effect of the density and distribution of water retention points on the performance of the PC-PTF, six scenarios were considered (Fig. 2). In each scenario, a unique set of measured water retention points from the dry and/or wet part of the WRC was provided in the training phase. The words 'dry' and 'wet' in this context explain the relative position of each water retention point among other measured points. This application does not necessarily match the scientific definition of the 'dry' and 'wet' parts of WRC. In scenario 1, all water retention points participated in the training phase, whereas every other point was picked up in scenario 2. In scenarios 3 and 4 dominant points were selected from wet parts and dry parts of WRC, respectively, while in scenarios 5 and 6, density of the points were increased in the intermediate and extreme matric potentials, respectively. The scenarios were not identical, though relatively similar, between the two datasets because of differences in available measured water retention points. The test samples for all scenarios were identical and contained all of the data points. The PC_{NN}PTF was derived in all of the scenarios but the PC_{SVM}PTF was just established in the first scenario for comparison with the equivalent PC_{NN}PTF. Log (h) = 6.9, with h matric head in cm, was assumed as the point with water content of zero in all soil samples and an extra point was added to the training set of all of PTFs in all scenarios to avoid negative prediction by PTFs moving towards the extremely dry part of WRC.

Table 1
Summary statistics of the datasets with data from Turkish soils (N = 135) and Belgian soils (N = 69).

	Turkey				Belgium			
	Max	Min	Mean	SD	Max	Min	Mean	SD
Clay (%)	62.2	8.0	34.4	13.6	64.0	1.0	14.2	12.7
Silt (%)	57.6	5.2	30.4	8.2	79.0	1.0	31.1	22.5
Sand (%)	83.6	5.9	35.1	1.6	97.0	4.0	54.7	29.6
BD (g cm^{-3})	1.66	0.93	1.25	0.17	1.77	1.05	1.47	0.15
OM (%)	3.85	0.01	1.24	0.65	6.71	0.09	1.54	1.49

BD: bulk density, OM: organic matter content, SD: standard deviation.

2.6. Evaluation method

The performance of the PTFs needs to be precisely evaluated. The PC-PTF in this study was desired to predict the water content continuously over a wide range of matric potentials rather than only predicting water content at the limited available points in the training phase. Therefore, a visual filtering process followed by a routine statistics-based evaluation was designed to fulfill this goal. In the visual filtering process, some samples from the cross-validation dataset were randomly selected and their WRCs were obtained using the RETC program version 6.02 (van Genuchten et al., 2009). It was assumed that the fitted WRC was fairly close to the real WRC of soil. Visual appearance of the predicted curves in comparison with the fitted ones was considered as the filtering criteria. The PTFs with unacceptable shapes were filtered in advance while statistics were calculated for the remaining models in order to recognize the best PTF.

In this study, the root mean square of error (RMSE) was selected as the main statistic to assess the performances of the PTFs. The correlation coefficient (r) and the mean bias error (MBE) were calculated as additional statistical parameters allowing recognition of any possible poor correlation and/or trend to over (under)estimation, respectively.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(E_i - M_i)^2}{n}} \tag{2}$$

$$r = \frac{\sum_{i=1}^{n}(M_i - \overline{M})(E_i - \overline{E})}{\sqrt{\sum_{i=1}^{n}(M_i - \overline{M})^2 \sum_{i=1}^{n}(E_i - \overline{E})^2}} \tag{3}$$

$$\mathrm{MBE} = \sum_{i=1}^{n}\frac{E_i - M_i}{n} \tag{4}$$

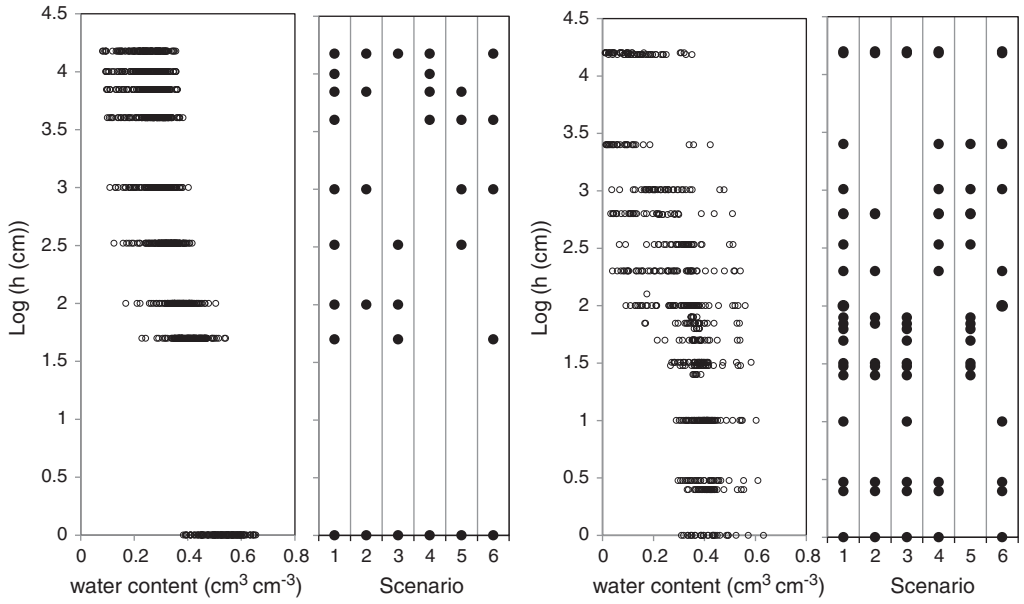

Fig. 2. Density of the available water retention points in the training phase of the scenarios with the datasets from Turkey (left panel) and from Belgium (right panel).

where M is the measured water content, E is the estimated water content, $\overline{E}$ is the average of estimated values, $\overline{M}$ is the average of measured values and n is the number of data (i.e. water retention points). The program IRENE (www.isci.it/tools) was used for identifying the statistics. In order to involve all of the samples in the test part and in turn achieving reliable results, and to determine the uncertainty of the calculations, the modeling steps were repeated five times (Haghverdi et al., 2014). The datasets were randomly divided into five equal portions, and modeling was repeated five times such that a different portion was assigned to the test dataset each time. Finally, the statistics were individually calculated for each of the five test portions where the standard deviation of the results was considered as uncertainty of the models.

3. Result and discussion

3.1. SVM versus NN

Eleven processing elements for both datasets were employed to optimize the PC_{NN}PTF. This number was chosen based on error variation in cross-validation data while the better result in the training phase was observed in a model with the maximum amount of neurons in the hidden layer, i.e. 15. This confirms the necessity of considering a cross-validation set or other available methods to avoid overtraining in NN-based PTFs.

The optimal parameters to establish the most accurate PC_{SVM}PTF, were $C = 35$, $\varepsilon = 0.05$, $\gamma = 6.3$ for the dataset with Belgian soils and $C = 43$, $\varepsilon = 0.05$, $\gamma = 5.7$ for that with Turkish soils. Lamorski et al. (2008) reported relative insensitivity in the performance of SVM PTF to the parameter selection. Similar results were found in this study where RMSE of the PC_{SVM}PTFs remained almost constant for a wide range of parameters ε, C and γ. However, when WRC shape was considered as a model selection criteria, the majority of models with satisfactory RMSE were filtered due to unacceptable curve shape. The accepted routine approach to select the best PTF is to compare the performance of the candidates by estimating some statistics over the test dataset. This process, however, appeared to be misleading and insufficient in the case of the PC-PTF (especially SVM-based PTF). The initial attempts for finding the best SVM model revealed that low values of RMSE in training and even cross-validation samples may belong to models that predict a linear matric potential–water content relationship rather than a nonlinear one. Indeed, the real ability of the PC-PTF to interpolate/extrapolate the water content for new points could not be reflected adequately and, most likely, is masked within the calculated statistics.

Table 2 and Fig. 3 present the results of the modeling with SVM and NN based PTFs. Scattering of the data in Fig. 3 clearly illustrates the better performance of the PC_{NN}PTFs than the PC_{SVM}PTFs. There is a 0.007 $cm^3\ cm^{-3}$ difference between the RMSE of SVM and NN for the dataset from Turkey, while this difference even increases to 0.029 $cm^3\ cm^{-3}$ for the Belgian soil dataset. According to Vereecken et al. (2010), these RMSE values are between the most typical RMSE values for parametric PTFs predicting water retention, which typically vary between 0.034 and 0.085 $cm^3\ cm^{-3}$. MBE values in Table 2 and also Fig. 3 show slight tendencies to underestimation in PC_{SVM}PTFs for the Belgium dataset. In contrast to our findings, Twarakavi et al. (2009) showed that SVM outperformed NN in parametric PTF by 0.012 $cm^3\ cm^{-3}$ difference, on average. They adopted the same dataset as used for developing Rosetta, which is a NN-based parametric PTF, to derive SVM-PTFs but employed a different output structure. They introduced either supremacy of the SVM over the NN or different output structure of the used PTFs as a potential reason for the better result of the SVM over NN. Although Lamorski et al. (2008) reported better performance of SVM over NN, the average of RMSE values for the SVM based PTF was only 0.001 $cm^3\ cm^{-3}$ less than that for the NN based PTF. It should be noted here that Lamorski et al. (2008)considered point PTFs, Twarakavi et al. (2009) parametric PTFs, whereas in our study, a pseudo continuous PTF was examined. The better performance of either of DM methods could be related to the PTF type and database characteristics rather than to inherent supremacy of either of the DM methods.

Table 2
Performance evaluation of the PC_{NN}PTF and the PC_{SVM}PTF for Turkish and Belgian datasets.

Dataset	PTF	RMSE	r	MBE
Turkey	PC_{SVM}	0.054 ± 0.005	0.87 ± 0.023	−0.011/0.0169
	PC_{NN}	0.047 ± 0.009	0.91 ± 0.036	−0.004/0.016
Belgium	PC_{SVM}	0.069 ± 0.016	0.86 ± 0.07	−0.037/−0.005
	PC_{NN}	0.040 ± 0.009	0.96 ± 0.017	−0.009/0.012

* The reported RMSE and r values are the average and standard deviation over the five portions in the test phase. The reported MBE values are the min and max of the observed values over the five portions in the test phase.

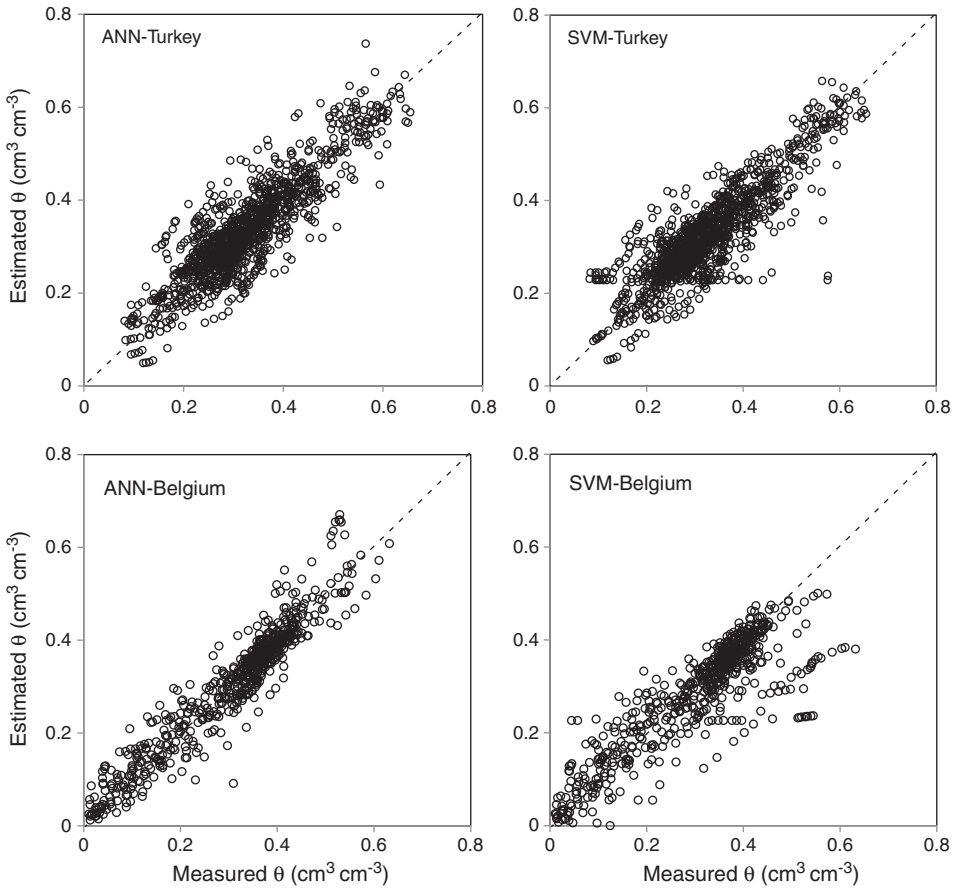

Fig. 3. Visual comparison of the performances of the PC_{SVM}PTF and the PC_{NN}PTF.

3.2. Impact of the measured points number and distribution

Fig. 4 shows the RMSE values of the PC_{NN}PTFs for the Belgian and Turkish datasets. The uncertainty of the performances was considered as the standard deviation of the PTF outputs through the five test portions. Overall, in both datasets the error increased moving towards scenario 5 but reduced at scenario 6. The better performance with the Belgian dataset occurred in scenario 1, whereas within the Turkish dataset the accuracy of the PTFs in scenarios 2, 3, 6 and 1 were about the same. Figs. 5 and 6 represent the WRCs of a randomly selected soil from the dominant texture in the Belgian and Turkish datasets, respectively. The gap between the WRCs in the training phase and test phase for the Belgian dataset was increased in scenarios with unbalanced training data such as scenarios 3, 4 and 5 (Fig. 5) and scenario 5 for the Turkish dataset. This illustrates that NN was not able to find a generalized shape for the WRC in the absence of sufficient data in the training phase, although the model may still work fairly well over the training phase.

Failure of the PTFs to project the correct WRC could easily be distinguished in PTFs with uneven scattered measured points in the training dataset, i.e. scenarios 3, 4 and 5 in both illustrated samples (Figs. 5 and 6), whereas some differences also occur due to different soil types in the datasets. Accuracy of predictions for both soils (Figs. 5 and 6) were enhanced when all water retention points (scenario 1), every other point (scenario 2) and extreme potentials (scenario 6) were utilized in the PTF development process. Almost identical curves for scenarios 1 and 2 for the coarse soils proved that accurate estimation is possible with less water retention points but covering the entire

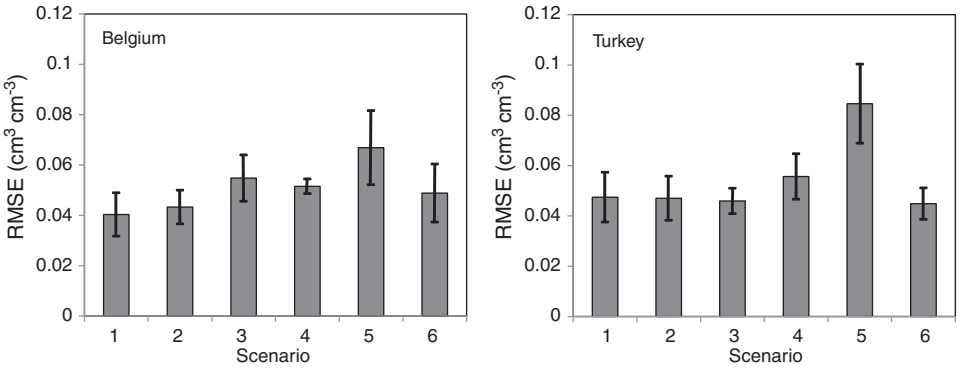

Fig. 4. RMSEs for the scenarios and datasets on testing phase.

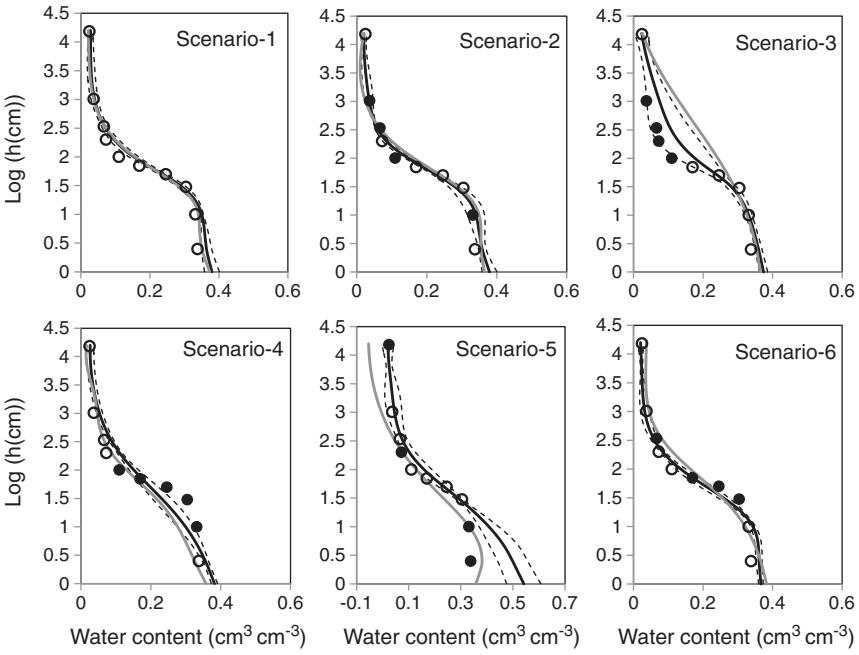

Fig. 5. Performance of the $PC_{NN}PTF$ for a random coarse soil in the Belgian dataset in all of the scenarios. The bold gray line shows the performance of the derived PTF in the test phase; the bold and dashed black lines show the average performance of the PTF in the training phase and uncertainty (i.e. the average ± standard deviation), respectively. The circles show the measured water retention points where the black circles were absent in the training phase.

range of the WRC (Fig. 5). However, the density of the measured points plays a significant role for fine soils especially at the mid moisture ranges of the curve (Fig. 6, scenarios 1 and 2). PTFs in scenarios 3 and 4, which lacked some adequate points on the WRC in the training phase, showed poor estimation for both soils. For example, taking only one water retention point at the dry boundary ($\log h = 4.2$) (scenario 3) and only a few in the wet part (scenario 4) triggered an overestimation at the dry end and an underestimation at the wet end of a coarse-textured Belgian soil, respectively. However, both scenarios yielded a high accuracy in the rest of the curves due to the use of all available points in this ranges. Similarly, the same error trend was also observed for the fine-textured soils of the Turkish dataset in scenario 3 (Fig. 6), whereas using a limited number of points in the dry range resulted in lower estimation values for such soils. One of the worst WRCs was obtained in scenario 4 due to lack of using any point between the saturation and $\log h = 3.6$ but many in the dry boundary.

For the coarse-textured soil, the odd shape with negative water content prediction at the very dry part of the curve and underestimation at

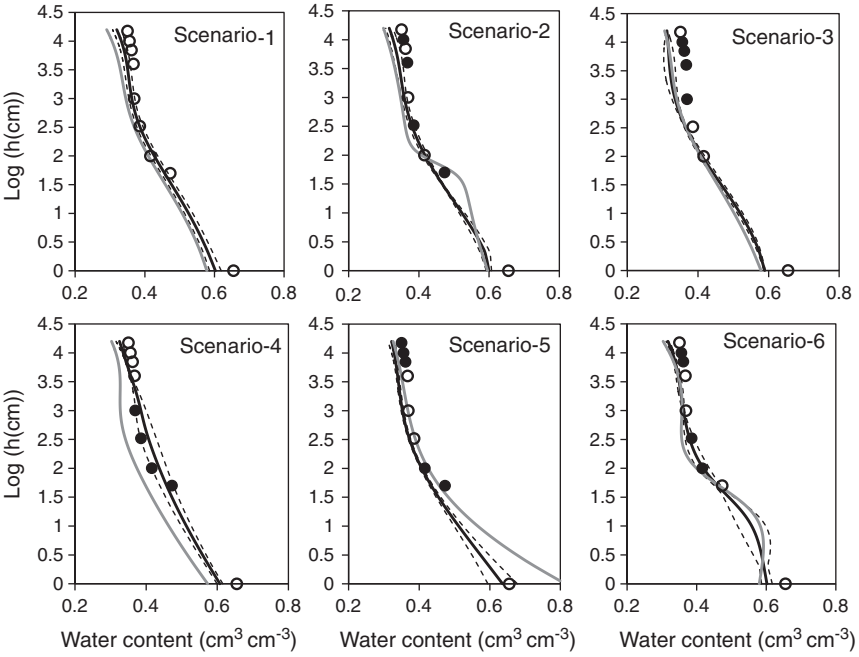

Fig. 6. Performance of the $PC_{NN}PTF$ for a random fine soil in the Turkish dataset in all of the scenarios. The bold gray line shows the performance of the derived PTF in the test phase; the bold and dashed black lines show the average performance of the PTF in the training phase and uncertainty (i.e. the average ± standard deviation), respectively. The circles show the measured water retention points where the black circles were absent in the training phase.

the very wet part in scenario 5 (Fig. 5) demonstrates that it is essential to have some points on the wet and dry boundaries in the training dataset. On the other hand, the absence of boundary points in scenario 5 for the fine-textured soil in Fig. 6 caused a very high prediction in the wet part of the curve and no negative prediction in the dry part.

Except for scenario 1, the number of the applied measured points in the training phase is almost the same in the rest of the scenarios. The variation in accuracy among scenarios emphasized the sensitivity of the PC_{NN}PTF to the density of the available points in the training dataset. This result is in agreement with the finding of Jain et al. (2004) who utilized NN for fitting WRC. They noted that application of NN requires the measured points over the entire range of the WRC. Furthermore, the promising performance of the PTFs in scenario 2 and somehow scenario 6 for coarse textured soils in our study showed that deriving accurate PTFs with a few measured points is possible if the available points cover the entire range efficiently and hence could picture the original shape of the WRC. If one wants to establish a parametric PTF, however, reducing the number of water retention points increase the risk of over-parameterization (Minasny and McBratney, 2002). This finding is important because establishing local soil databases is still an in-progress goal in developing countries, where decreasing the required measured water retention points will significantly reduce the associated cost and time of the lab/field works.

Fig. 7 illustrates the average performance of the PTFs for the different scenarios together with the percentage of the available data for each soil texture in the datasets. Overall, the accuracy of the PTFs was inversely proportional to the number of the available samples for each soil texture. For example, the highest error accrued for textures with the lowest amount of data which were fine and coarse-textured soils for the Belgian and Turkish datasets, respectively. Vereecken et al. (2010) reviewed the performance of parametric PTFs from different studies and showed that the percentage share of the textural classes in each dataset may affect the performance of the models. Texture-related parameters also affected the error to some extent. For instance, the error turned out to be relatively low for medium-textured soils, which were not well represented in both datasets. This could be related to either the shape of the WRC for medium-textured soils and in turn, the ease of modeling or to the correlation of the input attributes to water retention data.

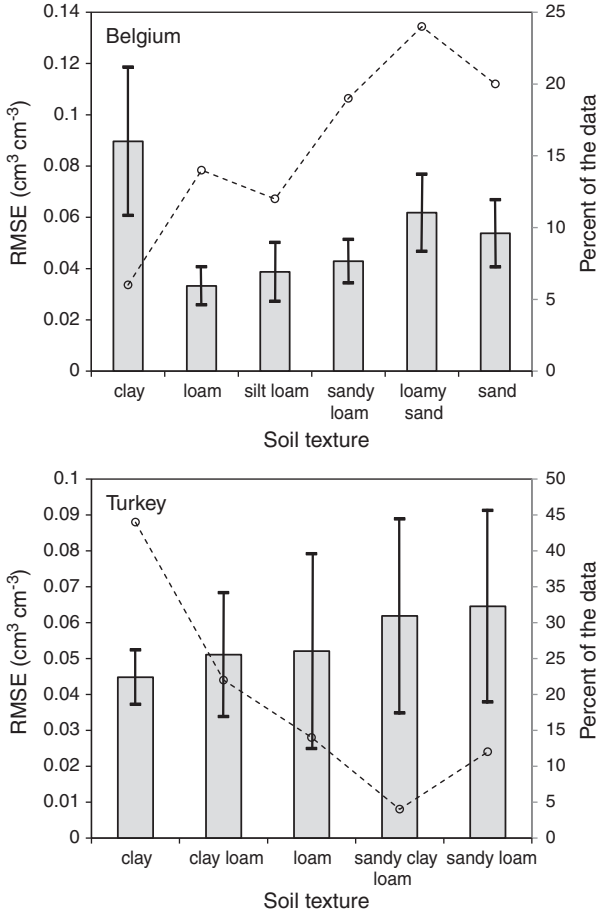

Fig. 7. Texture-based performance of the PTFs on Belgium dataset (top panel) and Turkish dataset (bottom panel). For each soil type the gray columns and the dashed lines/empty circles are RMSE and percentage of the data in datasets, respectively.

4. Conclusion

The PC-PTF was successfully implemented to predict WRC. The inherent supremacy of SVM-PTF over the NN-PTF approach as was observed in previous studies on PTF development, was not confirmed in this study, where it was shown that for two independent datasets, NN outperformed SVM. This study firmly proved the significance of data quality on the performance of the PC_{NN}PTF. A PC-PTF derived with a limited number of the measured points may still work fine as a point PTF, but it does not enable one to predict the whole WRC and, hence, should not be used. However, the number of measured points in the training phase could be decreased without significant side effects, i.e., drastic increase in error or false WRC shape, but only if the distribution of the measured water retention points is well enough to adequately represent the shape of the soil WRC. Optimizing the number of measured water retention points can significantly reduce the required time and the cost to establish a local database which is a prime limitation in most of the developing countries to derive local PTFs.

References

Blake, G.R., Hartge, K.H., 1986. Bulk density, In: Klute, A. (Ed.), Methods of Soil Analysis. Part 1, second ed. Agron. Monogr., 9. ASA and SSSA, Madison, WI, pp. 363–375.
Botula, Y.-D., Nemes, A., Mafuka, P., Van Ranst, E., Cornelis, W., 2013. Prediction of water retention of soils from the humid tropics by the non-parametric k-nearest neighbor approach. Vadose Zone J. 12 (2).
Cornelis, W.M., Ronsyn, J., Van Meirvenne, M., Hartmann, R., 2001. Evaluation of pedotransfer functions for predicting the soil moisture retention curve. Soil Sci. Soc. Am. J. 65, 638–648.
Cristianini, N., Shawe-Taylor, J., 2000. An Introduction to Support Vector Methods. Cambridge Univ, Press, Cambridge.
Demuth, H., Beale, M., 2000. Neural Network Toolbox. Mathworks Inc.
Gee, G.W., Bauder, J.W., 1986. Particle size analysis, In: Klute, A. (Ed.), Methods of Soil Analysis. Part 1, second ed. Agron. Monogr., 9. ASA and SSSA, Madison, WI, pp. 383–411.
Genuchten, Van, Th, M., Simunek, F., Leij, F.J., Sejna, M., 2009. The RETC code (version 6.02) for quantifying the hydraulic functions of unsaturated soils. (http://www.hydrus3d.com).
Haghverdi, A., Cornelis, W.M., Ghahraman, B., 2012. A pseudo-continuous neural network approach for developing water retention pedotransfer functions with limited data. J. Hydrol. 442, 46–54.
Haghverdi, A., Ghahraman, B., Leib, B.G., Pulido-Calvo, I., Kafi, M., Davary, K., Ashorun, B., 2014. Deriving data mining and regression based water-salinity production functions for spring wheat (Triticumaestivum). Comput. Electron. Agric. 101, 68–75.
Jackson, M.L., 1958. Soil Chemical Analysis. Prentice Hall Inc., Eaglewood Cliffs, N.J.
Jain, S.K., Singh, V.P., van Genuchten, M.T., 2004. Analysis of soil water retention data using artificial neural networks. J. Hydrol. Eng. 9 (5), 415–420.
Lamorski, K., Pachepsky, Y., Slawihski, C., Walczak, R.T., 2008. Using support vector machines to develop pedotransfer functions for water retention of soils in Poland. Soil Sci. Soc. Am. J. 72 (5), 1243–1247.
Minasny, B., McBratney, A., 2002. The method for fitting neural network parametric pedotransfer functions. Soil Sci. Soc. Am. J. 66 (2), 352–361.
Mucherino, A., Papajorgji, P.J., Pardalos, P.M., 2009. Data mining in agriculture, vol. 34. Springer.
Nemes, A., 2011. Databases of soil physical and hydraulic properties. Encycl. Agrophys. 194–199.
Nemes, A., Schaap, M.G., Leij, F.J., Wösten, J.H.M., 2001. Description of the unsaturated soil hydraulic database UNSODA version 2.0. J. Hydrol. 251 (3), 151–162.
Pachepsky, Ya.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.
Schaap, M.G., Bouten, W., 1996. Modeling water retention curves of sandy soils using neural networks. Water Resour. Res. 32 (10), 3033–3040.
Tempel, P., Batjes, N.H., van Engelen, V.W.P., 1996. IGBP-DIS soil data set for pedotransfer function development. Work. Pap. 96/05. ISRIC, Wageningen, the Netherlands.

Twarakavi, N.K.C., Simunek, J., Schaap, M.G., 2009. Development of pedotransfer functions for estimation of soil hydraulic parameters using support vector machines. Soil Sci. Soc. Am. J. 73 (5), 1443–1452.
Van Genuchten, M.T., 1980. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. Soil Sci. Soc. Am. J. 44 (5), 892–898.
Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. John Wiley & Sons, New York.
Vereecken, H., Weynants, M., Javaux, M., Pachepsky, Y., Schaap, M.G., van Genuchten, M.T., 2010. Using pedotransfer functions to estimate the van Genuchten-Mualem soil hydraulic properties: a review. Vadose Zone J. 9 (4), 795–820.
Wösten, J.H.M., Pachepsky, Y.A., Rawls, W.J., 2001. Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic characteristics. J. Hydrol. 251, 123–150.