Journal of Hydrology 442–443 (2012) 46–54

Contents lists available at SciVerse ScienceDirect

Journal of Hydrology

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

A pseudo-continuous neural network approach for developing water retention pedotransfer functions with limited data

A. Haghverdi [a,*], W.M. Cornelis [b], B. Ghahraman [a]

[a] Department of Water Engineering, Faculty of Agriculture, Ferdowsi University of Mashhad (FUM), P.O. Box 91775-1163, Mashhad, Iran
[b] Department of Soil Management, Ghent University, Coupure Links 653, B-9000 Ghent, Belgium

ARTICLE INFO

Article history:
Received 24 February 2011
Received in revised form 16 December 2011
Accepted 28 March 2012
Available online 11 April 2012
This manuscript was handled by Philippe Baveye, Editor-in-Chief, with the assistance of Pascal Boivin, Associate Editor

Keywords:
Pseudo-continuous PTF
Accuracy
Reliability
Water retention curve

SUMMARY

In this study, a new approach, which we called pseudo-continuous, to develop pedotransfer functions (PTFs) for predicting soil–water retention with an artificial neural network (ANN) was introduced and tested. It was compared with ANN PTFs developed using traditional point and parametric approaches. The pseudo-continuous approach has a continuous performance, i.e. it enables to predict water content at any desirable matric potential, but without the need to use a specific equation, such as the one by van Genuchten. Matric potential is considered as an input parameter, which enables to increase the number of samples in the training dataset with a factor equal to the number of matric potentials used to determine the water retention curve of the soil samples in the dataset. Generally, the pseudo-continuous functions performed slightly better than the point and parametric functions. The root mean square error (RMSE) of the pseudo-continuous functions when considering local data for training and testing, and with both bulk density and organic matter as extra input variables on top of sand, silt and clay content, was 0.027 $m^3\ m^{-3}$ compared to 0.029 $m^3\ m^{-3}$ for both the point and parametric PTF. The increased number of samples in the training phase and the selection of matric potential as an input variable enabling to predict water content at any desired matric potential are the most important reasons why pseudo-continuous functions would need more intention in the future. Uniformity in the training and test dataset was shown to be important in deriving PTFs. We finally recommend the use of pseudo-continuous PTFs for further improvement and development of PTFs, in particular when datasets are limited.

1. Introduction

Although nowadays soil hydraulic properties are amongst the most important parameters in agricultural research, in using irrigation and drainage models, and for studying water movement in the unsaturated zone of the soil, they are not readily available. Therefore, pedotransfer functions (PTFs) are developed to predict hydraulic properties from primary soil properties with a suitable mathematical relation. Extensive research in the past has focused on improving estimates of the hydraulic properties using PTFs (Vereecken et al., 2010).

Two main categories of methods for deriving PTFs can be distinguished: statistical regression techniques (linear and nonlinear models) and data mining and exploration techniques (e.g., artificial neural networks and group methods of data handling) (Vereecken et al., 2010). Recently some new mathematical methods, such as support vector machines (Lamorski et al., 2008; Twarakavi et al., 2009) and nonparametric nearest neighbor methods (Nemes et al., 2006a, 2006b and 2010), were used to predict water retention properties from basic soil properties. In general, methods based on artificial neural networks (ANNs) have led to PTFs that performed best in terms of basic performance indicators such as the root mean square error (RMSE) (Vereecken et al., 2010). This strong interest in using machine learning algorithms instead of traditional procedures such as multiple regressions for deriving soil hydraulic PTFs proves their ability to model the interaction of soil and water as a very complex system. The most important drawback of regression type PTFs became apparent when large databases were used for estimating hydraulic parameters (Wösten et al., 2001). In spite of the fact that ANNs are very powerful for deriving hydraulic PTFs, they are very data demanding and their application has only become possible when used together with a large database (Baker and Ellison, 2008). However, since the introduction of ANN derived PTFs in the mid 90s (Pachepsky et al., 1996), nobody answered how we can use this technique for deriving PTFs with limited data.

Tools such as Neuropack (Minasny and McBratney, 2002b) and Rosetta (Schaap, 2000) have been developed for deriving PTFs using ANNs. Neuropack, e.g., can be utilized to first fit pedotransfer functions using ANNs. The trained networks are subsequently used

* Corresponding author. Tel.: +98 5118455378, mobile: +98 9153023558.
E-mail addresses: amirhaghverdii@gmail.com (A. Haghverdi), wim.cornelis@UGent.be (W.M. Cornelis), bijangh@um.ac.ir (B. Ghahraman).

0022-1694/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jhydrol.2012.03.036

to validate and predict hydraulic properties of new soil samples. Neuropack consists of two programs: Neuropath and Neuroman. Neuropath is a general single layer neural network that can model any inputs-outputs relationship. Meanwhile, Neuroman is a neural network that predicts parametric PTFs. Both programs have a user-friendly interface with robust algorithms (Minasny and McBratney, 2002b). Neuropack has been used over a wide range of geographic areas including USA, Italy and Australia for developing PTFs (Minasny and McBratney, 2002a; Ungaro et al., 2005; Sharma et al., 2006). Minasny and McBratney (2002a) used Neuropack and compared its performance with Rosetta. They showed that Neuropack has better accuracy and less bias as compared with Rosetta.

The level of reliability of any given PTF is highly correlated to the specific composition of the calibration dataset, which in turn may reflect the geographic origin of the dataset. For this reason, the extrapolation of PTFs beyond their statistical training limits and their geographical training area should always be preceded by a careful evaluation of their applicability to specific datasets (Cornelis et al., 2001). On the other hand, in many countries and regions of the world, sufficient soil hydraulic data for deriving PTFs are lacking. Therefore, PTFs that are derived from origins different than those for which they were originally developed, are widely used. For identifying the level of influence of homogeneity in training and test datasets on the performance of PTFs, the ANN PTFs can be evaluated from two different aspects: testing (using the same dataset for training and testing, scenario 1) and validation (using different datasets for training and testing, scenario 2).

Generally two different kinds of ANN PTFs have been most frequently used by authors: point PTFs and parametric PTFs (respectively Type 2 and Type 3 PTFs in Wösten et al., 2001). Outputs from point PTFs are water contents at predefined potentials, which means that a continuous water retention curve at all matric potentials is not given. To obtain a continuous PTF that predicts water content at any matric potential, the coefficients of a closed-form analytical water retention equation need then to be determined by curve fitting. On the other hand, using a parametric PTF supposes that the relationship between water content and matric potential can be described adequately by a soil hydraulic model with a certain number of parameters, e.g., the Brooks and Corey (1964) or the van Genuchten (1980) equations. The main disadvantage of parametric PTF is that sometimes, the real shape of the water retention curves is not similar to the chosen equation shape for all soil samples. In addition, some problems are reported correlating the parameters of soil hydraulic equations to basic soil properties (Minasny and McBratney, 2002b). Furthermore, parametric PTFs predetermine which equation the user is to use, which is for most PTFs or the van Genuchten or the Brooks and Corey equation. We, therefore, introduced a new method for deriving PTFs that have a (pseudo) continuous performance, but without the need to use a specific equation. In addition, the special topology of this PTF enables the user to apply it with limited data information.

The objectives of this study were, (1) to introduce and evaluate a new kind of PTF which we call pseudo-continuous PTF, (2) to evaluate the accuracy and reliability of the derived PTF, and (3) to compare its performance with that of PTFs developed on the same dataset using the point and parametric approaches.

2. Materials and methods

2.1. Soil samples

Three different datasets were used in this study. Table 1 and Fig. 1 show the physical characteristics and the scattering of soil samples in the soil texture triangle, respectively. The first dataset, DS1, contains 122 soil samples from northeastern (Haghverdi et al., 2010) and northern Iran (Khoshnood Yazdi and Ghahraman, 2004). From the northern site, 50 disturbed and undisturbed (226 cm^3) soil samples were collected from the surface soil (0–30 cm). Bulk density and hydraulic properties were identified using undisturbed samples while the rest of properties were measured using disturbed samples. Soil sampling was done according to a quadrangle grid with 200 m node spacing. Particle-size distribution was determined by the hydrometer method (Gee and Bauder, 1986). Organic carbon content (OC) and bulk density (BD) were determined by the Walkley and Black method (USDA, 1982) and using the soil clods method described by Blake and Hartge (2002), respectively. Water content of the samples was measured at −5, −33, −100, −500, −1500 kPa imposed in a pressure plate apparatus (Soilmoisture Equipment, Santa Barbara CA, USA). From the northeastern site, 72 disturbed and undisturbed (180 cm^3) soil samples were collected during another independent study. Samples were collected from random locations within the site. Particle-size distribution, organic matter content and bulk density was determined as above. Water contents of those samples were measured at −33, −100, −400, −700, −1000 and −1500 kPa imposed in a similar pressure plate apparatus. The second and third dataset, DS2 and DS3, were established from Australian soils and are provided in the Neuropack software package. DS2 contains 622 soils with water contents measured at 0, −5, −30, −500 and −1500 kPa. In DS3, there are 150 soil samples having information on water content at many matric potentials, more than 15 points which were not identical in all samples. The DS2 and DS3 contain similar soil basic properties as DS1, except OC information for DS3.

2.2. Pedotransfer functions

We ran three different PTFs in this study, i.e. point, parametric and pseudo-continuous PTFs. The typical topologies of point, parametric and pseudo-continuous neural network PTFs used in this study are presented in Fig. 2. In case of *point* PTFs, water contents at specific matric potentials were predicted according to the common information between training and test data using the Neuropath software. Neuropath attempts to find such relationships by adjusting the weights through the process of training. An optimization procedure using the NL2SOL adaptive nonlinear least squares algorithm (Eq. (1)) was applied for training. The objective is to minimize the sum of squares of the residuals between the measured and predicted outputs:

$$O(W,U)=\sum_{i=1}^{N_s}\sum_{k=1}^{N_o}\left(\widehat{P}_{ik}(x_i)-P_{ik}\right)^2 \tag{1}$$

where N_s is the number of samples, N_o is the number of outputs, W and U are the weights of the hidden and output layer, respectively, P is the measured output, and $\widehat{P}$ is the predicted output from inputs x.

To derive *parametric* PTFs, the van Genuchten equation, as the most widely used soil hydraulic model, was chosen. It is among the best performing water retention models (Cornelis et al., 2005), except when describing the complete water retention curve between saturation and oven dryness (Khlosi et al., 2008). It is written as:

$$\theta(h)=\theta_r+\frac{\theta_s-\theta_r}{(1+|\alpha\psi|^n)^m} \tag{2}$$

where θ_r and θ_s are the residual and saturated water content, respectively, α is the scaling parameter, n is the curve shape factor, m is an empirical constant, which can be related to n as $m = 1-1/n$, and ψ is matric potential. The coefficients of the van Genuchten equation of DS1 were achieved from an optimization process with the RETC program version 6.02 (van Genuchten et al., 2009). Since

Table 1
Physical characteristics of soil samples.

	DS1				DS2				DS3			
	Max	Min	Mean	SD	Max	Min	Mean	SD	Max	Min	Mean	SD
Sand (%)	77.2	10.2	38.8	12.8	98.9	0.0	43.9	30.6	95.5	2.0	52.6	28.6
Silt (%)	59.6	14.6	38.9	9.9	72.7	1.1	35.0	16.5	68.3	1.0	28.5	18.8
Clay (%)	56.0	8.2	22.3	7.9	81.7	0.0	21.1	19.4	64.0	0.0	18.8	12.9
OC (%)	1.9	0.1	0.7	0.4	25.6	0.0	0.9	2.0				
BD (g cm^{-3})	1.63	1.37	1.40	0.04	1.87	0.57	1.46	0.19	1.70	1.08	1.45	0.13

*SD: standard deviation, BD: bulk density, OC: organic carbon.

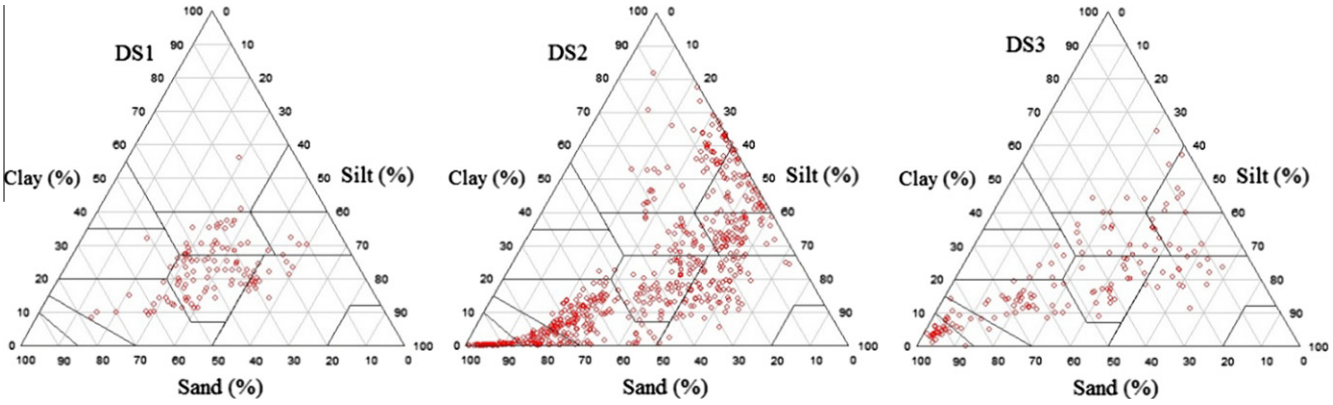

Fig. 1. Soil texture of the three datasets, DS1, DS2 and DS3, used in this study. Clay corresponds to 0–2 μm, silt to 2–50 μm and sand to 50–2000 μm.

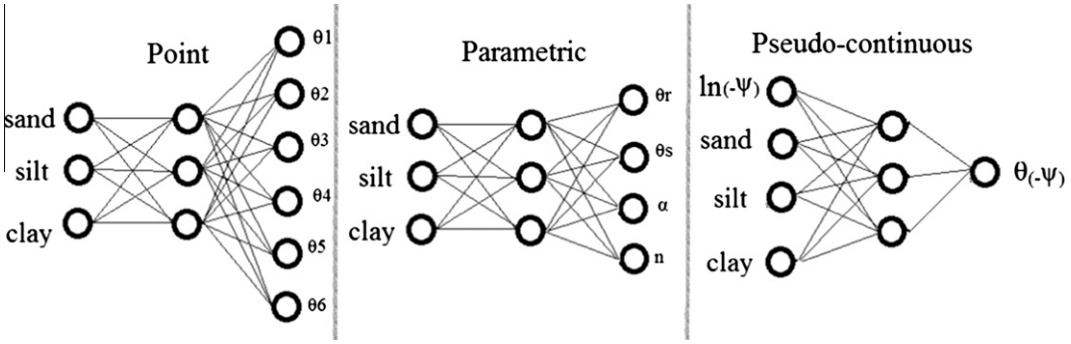

Fig. 2. The typical topologies of the point, parametric and pseudo-continuous neural network PTFs used in this study. Sand, silt and clay are sand, silt and clay percentages which are the common input predictors of PTFs. Bulk density and organic matter content can be added as optional extra predictors. $\theta_1, \theta_2, \ldots, \theta_6$ are volumetric water contents which are the outputs of the point PTF when using a dataset containing six points of the water retention curve for each sample. θ_r, θ_s, α and n are the parameters of the van Genuchten equation (Eq. (2)) which in turn are the outputs of the parametric PTF. ln $(-\psi)$ is matric potential which is the extra input predictor of the pseudo-continuous PTF. $\theta(-\psi)$ is the volumetric water content at $-\psi$ matric potential which is the output predictor of the pseudo-continuous PTF. Different $-\psi$ values yield different water contents.

some of the parameters exhibit non-normal distributions, they were transformed, e.g., $(\theta_r)^{1/2}$, $\ln(\alpha)$, and $\ln(n-1)$. A modified objective function for neural network training, proposed by Minasny and McBratney (2002a), was used to derive parametric PTFs. This objective function minimizes the difference between the measured water content and the one calculated from the predicted parameters in a secondary training process after doing an initial routine training step to estimate van Genuchten equation parameters. The Neuroman software was used to derive parametric PTFs.

To develop a *pseudo-continuous* PTF, a new topology for neural network PTFs was introduced. In this method, we consider the natural logarithm of matric potential as an input parameter, enabling the user to derive water content at any desired matric potential. Consequently, there is only one output parameter, θ, which shows the water content at the predefined matric potential that is considered as an input parameter. When using a wide range of matric potentials as input, a corresponding range of water contents will be predicted, and a (pseudo) continuous curve is obtained. The pseudo-continuous PTF was derived using the Neuropath software, using the same optimization procedure as with the point PTF.

The bootstrap method (Breiman, 1996) was used to establish more reliable PTFs and for avoiding local minima, the training process was iterated. The number of bootstraps and iterations were 50 and 100, respectively. Both the resampling procedure and the iteration of training were done by the Neuropack software package, while the optimized weights after training used in the testing part and the performance of the PTFs in terms of RMSE based on the test dataset were used to rank the models.

2.3. Accuracy and reliability

In this study, two different scenarios for training ANN were applied for testing the accuracy and reliability of PTFs as defined in

Vereecken et al. (2010). In scenario 1, for identifying the accuracy of the PTFs, we used DS1 for both the training and testing process. Out of 122 existing data in DS1, 80% were selected for training and the remaining 20% were selected for testing randomly. In scenario 2, for evaluating the reliability of the PTFs, DS2 and DS3 were chosen as training datasets for deriving point and parametric PTFs, respectively, whereas both DS2 and DS3 were used to derive pseudo-continuous PTFs. The complete set of 122 samples in DS1 was used for the testing part of those three PTF approaches. Since DS2 only contains samples with only five water retention data pairs, it is not appropriate for estimating parameters of the four-parameter van Genuchten equation because of lack of degrees of freedom whilst DS3 is not appropriate for deriving point PTF due to the high and unequal number of water content information across the samples. However, for making a fair comparison between the reliability of PTFs and because of the flexible structure of pseudo-continuous PTFs, both DS2 and DS3 were used to derive pseudo-continuous PTFs.

It should further be emphasized that the difference in sampling strategy and sample size of the two datasets used to establish DS1 does not affect the results, because identical samples were used to test and validate PTFs. Table 2 shows the different combinations of input and output variables of the PTFs that were evaluated in this research.

2.4. Performance criteria

Choosing the most suitable statistics for evaluation of PTFs is subject of dispute. Generally, the combination of selected statistics should picture the scattering of data around the 1:1 line. Even the root mean square of error (RMSE), as the most popular statistic, cannot show this aspect of the scatter plot alone. For instance, in the case of overestimation (underestimation) or very poor correlation between measured and predicted data, using RMSE alone is questionable. In this study, RMSE was selected as the main statistic for evaluating and ranking the models, whilst additionally the correlation coefficient (r) and the mean bias error (MBE) were considered to evaluate whether there is a correlation and/or overestimation (underestimation) problem in any PTF. There are two important reasons for ignoring r and MBE in the evaluation process. Firstly, r and MBE are not suitable statistics for independent use to evaluate PTFs, since the information they provide does not reflect the PTFs accuracy. In addition, considering more than one statistic for evaluating the PTFs can be confusing and, especially when they do not show the same behavior, the judgment will be very complicated. The above performance indicators can be defined as:

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

$$\mathrm{RMSE} = \left[\frac{1}{n}\sum_{i=1}^{n}(E_i - M_i)^2\right]^{1/2} \quad (4)$$

$$\mathrm{MBE} = \frac{1}{n}\sum_{i=1}^{n}(E_i - M_i) \quad (5)$$

where M is the measured value, E is the estimated value, $\overline{M}$ is the average of measured values, $\overline{E}$ is the average of estimated values and n is number of data for each soil sample. The program IRENE (http://www.isci.it/tools), a data analysis tool designed to provide easy access to model evaluation techniques, was used for identifying the statistics. The program provides extensive statistical capabilities with tools for a variety of needs.

3. Results

The summary of statistics and the scatter plots of measured versus estimated water contents for the test datasets are shown in Table 3 and Fig. 3. Both scenarios showed reasonable accuracy in predicting water content. MBE and r varied from 0.004 to 0.015 (in absolute terms) and 0.81 to 0.95, respectively, which proves there was not even a single case of substantial overestimation or underestimation, or poor correlation between measured and estimated data.

Generally, the best results were achieved with the pseudo-continuous PTFs and, as could be expected, with scenario 1, for all input variable combinations, with RMSE ranging between 0.027 and 0.029. In case of scenario 2, RMSE of the pseudo-continuous PTF varied from 0.036 to 0.037. RMSE of the conventional point and parametric PTFs varied for the scenarios 1 and 2, respectively, from 0.029 to 0.033 and from 0.032 to 0.053. As said before, it should be noted that in scenario 2, we used DS2 for training the point PTF and DS3 for training the parametric PTF. To avoid obscuring the results, we did not compare the reliability of the parametric PTFs with that of the point PTFs.

Unequal number of samples and lack of data at some matric potentials with the point PTFs hamper the comparison between the different methods. To enable a more complete comparison, Table 4 shows the performance of the PTFs at each individual matric potential.

4. Discussion

4.1. Importance of input variables

According to Table 3, it seems that adding both OC and BD as extra predictors only slightly improved the overall accuracy and reliability of all PTFs, although the observed differences were statistically not significant at the 0.95 probability level. When considering RMSE per matric potential (Table 4), it can be seen that

Table 2
Combinations of input and output parameters of scenarios 1 and 2 for PTFs.[a]

PTF	Scenario	Train dataset	Test dataset	Inputs	Outputs
Point	1	DS1 (80%)	DS1 (20%)	SSC, BD, OC	$\theta_{(-\psi)}$
Point	2	DS2	DS1 (100%)	SSC, BD, OC	$\theta_{(-\psi)}$
Parametric	1	DS1 (80%)	DS1 (20%)	SSC, BD, OC, $\ln(-\psi)$[b]	θ[b]
Parametric	2	DS3	DS1 (100%)	SSC, BD, $\ln(-\psi)$	θ
Pseudo-continuous	1	DS1 (80%)	DS1 (20%)	SSC, BD, OC, $\ln(-\psi)$	θ
Pseudo-continuous 2	2	DS2	DS1 (100%)	SSC, BD, OC, $\ln(-\psi)$	θ
Pseudo-continuous 3	2	DS3	DS1 (100%)	SSC, BD, $\ln(-\psi)$	θ

[a] SSC: particle size distribution containing sand, silt and clay percentages; BD: bulk density, OC: organic carbon; θ: water content as a single output parameter related to a single matric potential $\ln(-\psi)$ as an input parameter; $\theta_{(-\psi)}$: volumetric water content at $-\psi$ matric potential (kPa).
[b] The parametric PTFs first estimate the van Genuchten parameters, which were then used to estimate water content at the matric potential of interest.

Table 3
Performance indicators of modeling with point, parametric and pseudo-continuous PTFs.

	Scenario 1			Scenario 2		
	RMSE	MBE	*r*	RMSE	MBE	*r*
Point						
SSC	0.031	0.004	0.90	0.035	−0.008	0.95
SSC, BD	0.033	0.013	0.91	0.033	−0.008	0.95
SSC, OC	0.029	0.011	0.93	0.033	−0.009	0.95
SSC, BD, OC	0.030	0.010	0.92	0.032	−0.008	0.95
Parametric						
SSC	0.032	0.009	0.94	0.046	0.010	0.86
SSC, BD	0.032	0.011	0.94	0.053	0.015	0.81
SSC, OC	0.030	0.011	0.95			
SSC, BD, OC	0.029	0.009	0.95			
Pseudo-continuous DS2						
SSC	0.029	0.007	0.95	0.037	−0.014	0.91
SSC, BD	0.028	0.008	0.95	0.037	−0.014	0.92
SSC, OC	0.028	0.008	0.95	0.036	−0.014	0.92
SSC, BD, OC	0.027	0.006	0.95	0.036	−0.014	0.92
Pseudo-continuous DS3						
SSC				0.044	0.009	0.88
SSC, BD				0.049	0.010	0.84

increasing the number of predictors did, however, not always improve the performance of the PTFs. In scenario 2, adding extra input predictors to pseudo-continuous DS2 (which used DS2 for training), improved the performance, whereas it did not to pseudo-continuous DS3 (which used DS3 for training). In a similar way, the performance of the parametric PTFs in scenario 2, which used DS3 for training, was not improved by adding BD. This fact reveals that the efficiency of extra input predictors can be affected by the characteristics of the training and testing dataset. Many authors (e.g., Schaap and Leij, 1998a; Minasny and McBratney, 2002a; Nemes et al., 2003; Patil and Rajput, 2009; Vereecken et al., 2010) reported that considering more input variables can improve the efficiency of ANN models. Our results support, however, the findings of Nemes et al. (2006a) and Manyame et al. (2007), who found that increasing the number of input variables does not necessarily lead to much better predictions.

4.2. Comparing scenarios

Generally, the PTFs in the scenario 1, though based on a much smaller training dataset, performed better than in scenario 2 (Tables 3 and 4). The difference between the two scenarios was significant at the 0.95 probability level for the parametric and pseudo-continuous PTFs. According to Schaap and Leij (1998b), using a similar dataset for training and testing results in a better performing PTF than using a different dataset for deriving PTFs. They developed ANN PTFs using the same algorithm and the same predictors, and validated them on data from three independent databases. They found that PTFs derived from one database gave systematically different estimations compared to PTFs derived from the other two datasets. Those estimations were improved somewhat when PTFs were derived from all available data. They concluded that the performance of a PTF depends on both the derivation

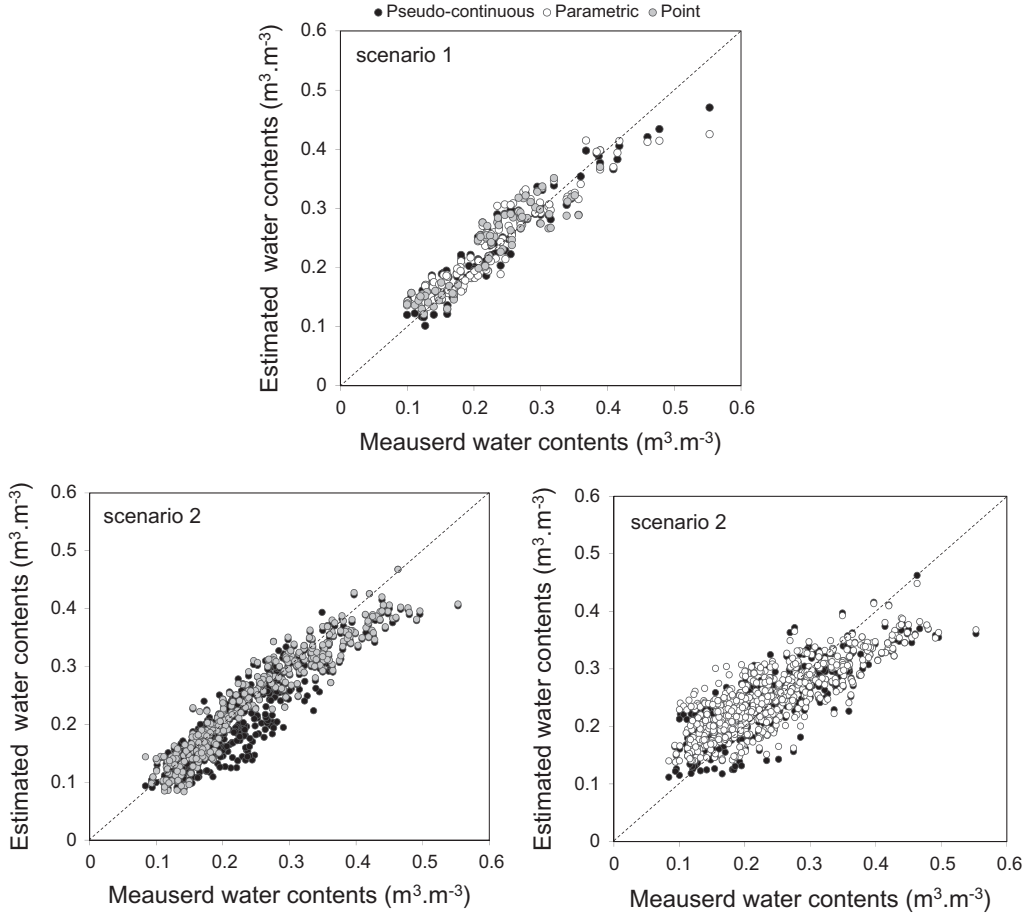

Fig. 3. Scatter plots of measured versus predicted water contents. The combination of input predictors in scenario 1 is SSC, BD and OC for all PTFs. The combination of input predictors is SSC, BD and OC in scenario 2 in point and pseudo-continuous DS2 PTFs, and SSC and BD in scenario 2 in parametric and pseudo-continuous DS3 PTFs.

Table 4
RMSE for each PTF at different matric potentials.

			Matric potential (kPa)							
PTFs	Scenario		−5	−33	−100	−400	−500	−700	−1000	−1500
Point	1	SSC		0.038	0.029					0.025
		SSC, BD		0.035	0.037					0.025
		SSC, OC		0.032	0.031					0.023
		SSC, BD, OC		0.034	0.031					0.024
Point	2	SSC	0.060	0.033			0.024			0.025
		SSC, BD	0.055	0.032			0.022			0.025
		SSC, OC	0.056	0.033			0.022			0.024
		SSC, BD, OC	0.053	0.031			0.020			0.025
Parametric	1	SSC	0.051	0.036	0.030	0.029	0.021	0.028	0.030	0.028
		SSC, BD	0.055	0.035	0.032	0.022	0.020	0.026	0.029	0.028
		SSC, OC	0.054	0.032	0.026	0.027	0.020	0.026	0.028	0.026
		SSC, BD, OC	0.050	0.033	0.027	0.025	0.018	0.024	0.026	0.025
Parametric	2	SSC	0.058	0.041	0.037	0.038	0.037	0.047	0.056	0.055
		SSC, BD	0.075	0.041	0.038	0.038	0.051	0.054	0.063	0.066
Pseudo-continuous	1	SSC	0.044	0.034	0.031	0.031	0.022	0.024	0.024	0.024
		SSC, BD	0.040	0.033	0.033	0.033	0.024	0.022	0.022	0.021
		SSC, OC	0.041	0.031	0.030	0.030	0.019	0.026	0.025	0.022
		SSC, BD, OC	0.036	0.032	0.030	0.030	0.020	0.024	0.023	0.021
Pseudo-continuous DS2	2	SSC	0.062	0.035	0.036	0.053	0.023	0.026	0.025	0.025
		SSC, BD	0.057	0.034	0.036	0.055	0.022	0.026	0.026	0.024
		SSC, OC	0.058	0.035	0.035	0.053	0.022	0.024	0.024	0.023
		SSC, BD, OC	0.056	0.033	0.036	0.055	0.021	0.025	0.025	0.024
Pseudo-continuous DS3	2	SSC	0.060	0.042	0.035	0.034	0.039	0.046	0.052	0.048
		SSC, BD	0.078	0.044	0.036	0.034	0.048	0.050	0.056	0.053

and evaluation datasets, as well as on the origin, size, and other characteristics. Applying scenario 1, and thus using samples from one identical dataset for training and testing, improved its performance (in terms of RMSE) on average with only 8% when establishing point PTFs. The average relative improvements in RMSE of the parametric and pseudo-continuous PTFs in scenario 1 over parametric, pseudo-continuous DS2 and pseudo-continuous DS3 were 38%, 23% and 40%, respectively. The largest differences between the scenarios can be observed at −400 kPa for the pseudo-continuous DS2 PTF and between −400 and −1500 kPa for the parametric and pseudo-continuous DS3 one. Various authors offered different explanations for the fact that using local data improves the PTF performance. Nemes et al. (2003) and Parasuraman et al. (2006) emphasized that an ANN model trained even on a small set of relevant data, when available, is better than training an ANN model on a large but more general dataset. Nemes et al. (2006a) showed that the improvement when using local data is related to the actual soil properties and is not origin specific. Our findings confirm the importance of using local data. Adding extra input variables such as BD and OC did, as was discussed above, not significantly affect the PTFs. The better results when using scenario 1 might be associated with other soil characteristics than those that were considered as input variables in this study. There are lots of important characteristics such as clay mineralogy and salinity of soils that largely influence the soil water status, but which are not accounted for as input variables in nearly any PTFs (Nemes et al., 2003). Sharma et al. (2006) illustrated that characteristics such as vegetation and topography significantly affect soil hydrologic phenomena and properties. Whenever these characteristics, if not included as input variables, are similar among training and test datasets, as in case of our scenario 1, the performance of the PTF may be improved. On top of that, also inter-correlations between some input/output variables were different among training samples in scenario 1, DS1, and scenario 2, DS2 and DS3, which affected the PTF's performance. For instance the correlation coefficients of sand (silt) with water contents were 0.27 (0.02), 0.71 (0.48) and 0.64 (0.57) in DS1, DS2 and DS3, respectively. Furthermore the correlation of matric potential and water contents was stronger in DS1 ($r = 0.8$) compared with DS2 ($r = 0.36$) and DS3 ($r = 0.28$) which in turn shows the differences in the real shape of soil samples' WRC among two origins, Iran and Australia.

4.3. Performance of PTFs at different matric potentials

Ignoring the effect of different combinations of input predictors, the mean RMSE for all PTFs in both scenarios at different matric potentials is presented in Fig. 4. The lowest RMSE values can be observed at −500 kPa (Table 4). Except for the parametric and pseudo-continuous DS3 PTFs in scenario 2, RMSE values at matric potentials equal to or below (more negative) −500 kPa appear to be significantly lower than those equal to or higher (less negative) than −400 kPa. In particular, matric potentials of −5 and −33 kPa appear to show the highest RMSE values (Fig. 4). According to Nemes et al. (2006a), reasons for this are the larger values of water content at high matric potentials which obviously causes greater variation, and the fact that water contents in the capillary and air-entry region of the soil–water retention curve (Jury and Horton, 2004) are more influenced by soil structure and pore-size distribution (Or and Wraith, 2002), which are only indirectly accounted for in the PTFs by BD. Rather surprising was the relatively large RMSE value at −400 kPa observed for the pseudo-continuous DS2 PTFs in scenario 2. The probable reason for this anomaly may be the lack of information at this matric potential in the training dataset, DS2, which means that water contents were interpolated at that matric potential in the pseudo-continuous PTF. The relatively large RMSE values at matric potentials $\leqslant$−500 kPa observed with the parametric and pseudo-continuous DS3 PTFs in scenario 2 may be due to the uneven number of training soil samples at different matric potentials. Only 9.85% of all existing information in DS3, which we used for developing the parametric and pseudo-continuous DS3 PTFs, corresponds to matric potentials $\leqslant$−500 kPa. Furthermore, the difference in nature and characteristics of the soils in DS1 and DS3 is most probable responsible for the large deviations observed in scenario 2 in the adsorption region (Jury and Horton,

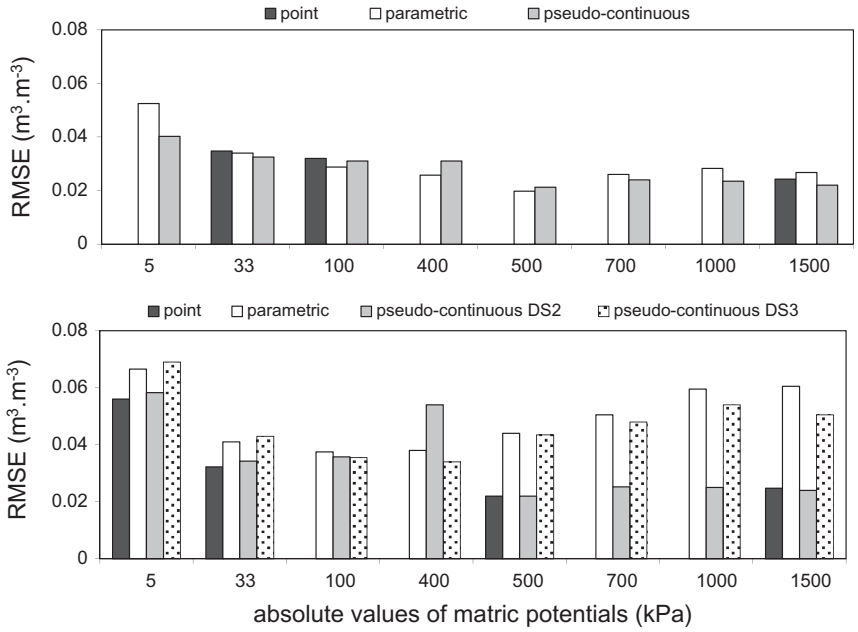

Fig. 4. Mean RMSE of all combinations of input predictors calculated with the different PTFs at different matric potentials for both scenario 1 (top panel) and scenario 2 (down panel).

2004), where adsorption forces and thus intrinsic soil characteristics are becoming most important.

A matric potential-dependent analysis of RMSE between saturation and permanent wilting point of various parametric PTFs by Vereecken et al. (2010) showed that the largest deviations occurred at intermediate matric potentials (approximately between −20 and −100 kPa) for PTFs using combinations of SSC, BD and OM as predictors. This could be in accordance with the higher deviations they observed in that range when considering the direct fits of the van Genuchten equation on the data used in developing the selected parametric PTFs. Since the pseudo-continuous approach is introduced in our study for the first time, these findings must be further investigated.

4.4. Comparing the performances of PTFs

The pseudo-continuous PTFs appear to perform slightly better than the parametric PTF (Tables 3 and 4). This demonstrates that using a predefined equation such as the one of van Genuchten is not necessary when wishing to derive PTFs using datasets with limited water retention data pairs. Artificial Intelligence methods to describe the water retention curve using observed data may result in better predictions of the real water retention curve than when using the van Genuchten equation. This finding confirms the result of Jain et al. (2004) who reported that neural networks can make equally or more accurate estimations of water retention compared with analytical hydraulic functions such as those by van Genuchten (1980) and Assouline et al. (1998). Although all of the PTFs in our study are based on a neural network, they are different in the way which they gain power from it. The pseudo-continuous approach is the only structure who's neural network tries to model the whole water retention curve. The fact that ANNs model the behavior of real-world data more accurate than the empirical methods, which is emphasized by Jain et al. (2004), is as well illustrated in our study. The RMSE values for our pseudo-continuous PTFs are also lower than those obtained by Twarakavi et al. (2009) when applying support vector machines (RMSE = 0.034) and ANN (using Rosetta) (RMSE = 0.044) for predicting water content at various matric potentials with the van Genuchten equation, although the difference in datasets among the studies does not allow us to make a firm conclusion. It should also be noted that in our test dataset, water contents at matric potentials below −400 kPa, which represent the dry part of WRC that is easier to estimate than the wet part, are greatly represented.

When comparing point and pseudo-continuous PTFs, the latter performs better than the point PTFs in 60% of the cases. When considering scenario 1, point PTFs show similar deviations than parametric ones. The general belief is that PTFs providing point estimates are more accurate than estimates based on parametric PTFs (e.g., Pachepsky et al., 1996; Tomasella et al., 2003; Vereecken et al., 2010). The possible explanation these authors give is that water contents at different matric potentials are generally affected by different soil properties and thus that using PTFs for each specific matric potential offers a better combination of these properties. Furthermore, parameterizations are often less appropriate in specific regions of the water retention curve, even when using a model like the one by van Genuchten, which is among the most flexible ones for describing the water retention curve between saturation and permanent wilting point (Cornelis et al., 2005).

Compared with the study of Nemes et al. (2006a), the pseudo-continuous approach showed lower RMSE values than the k nearest neighbor method at both field capacity (RMSE = 0.050) and permanent wilting point (RMSE = 0.035). However it should be noted that the error statistics would probably compare somewhat differently if all these methods were tested using the same data.

Apart from our promising results, the new topology of pseudo-continuous PTF enables the user to consider all the data at each matric potential in the training process. This is a very important advantage in comparison with the routine topology of point PTFs. The latter only allows estimating water content at matric potentials that are common between all samples of the training and test dataset. When the matric potentials at which water contents were determined are different or not completely overlapping between the training and test dataset, lots of data will remain useless. For example, in DS1, there are 122 soil samples available with water contents determined at five and six different matric potentials. Samples of DS1 were gathered from two different previous studies and the matric potentials at which water contents were measured

are not similar among them. With the point topology, the user can only model water content at −33, −100 and −1500 kPa, which are common among all 122 samples. With the pseudo-continuous approach, all data of each sample can be used.

Applying an ANN generally requires a huge dataset in the training phase. As discussed above, using different datasets from different locations for training and testing (scenario 2) almost always results in weaker predictions compared to training and testing on data from a unique location (scenario 1). Since local data are less bountiful, data conservation becomes a critical factor in ANN PTF construction (Baker and Ellison, 2008). Using pseudo-continuous PTFs can contribute to easing these problems. The important advantage of pseudo-continuous PTFs is their particular structure that enables to increase the number of instances used for training. Consider e.g., a soil sample with water content measured at five matric potentials. When one uses this sample to construct a point or parametric PTF, the output constitutes, respectively, five output parameters, each of them belonging to a specific neuron, or the coefficients of the used water retention equation. In pseudo-continuous PTFs, there is just one output neuron. For example, in DS1, there are 72 soil samples available with water contents at six matric potentials and 50 soil samples available with water contents at five matric potentials. When running a point or parametric PTF, one only has 122 rows for both training and testing phases. However, when using a pseudo-continuous approach, $(72 \times 6) + (50 \times 5) = 686$ rows of data will become available. The pseudo-continuous approach can thus be very useful for developing ANN-based PTFs when a limited number of soil samples are available. This is often the case in countries where a huge and strong database is not present, which constitutes most parts of the world. Although from practical point of view a pseudo-continuous neural network PTF is derivable even with a few samples, the user should not ignore the physical limitations of estimating the water retention curve using limited data. It should be emphasized that when using a pseudo-continuous approach for developing an ANN PTF, the complexity of the relationship among input-input and input–output has also increased compared with the traditionally derived ANN PTFs. The pseudo-continuous ANN PTF has to cover more nonlinearity and a more complex dependence of the output on the inputs, which in the traditional parametric approach is covered by the soil hydraulic equation, such as one by van Genuchten (1980). A sufficient amount of information for deriving a reliable and accurate PTF using the pseudo-continuous approach should be identified carefully in terms the number of samples, density of the measured points of the water retention curve, the number of input predictors and physical properties of soil samples.

5. Conclusions

In this study, three approaches for establishing PTFs have been applied and compared, i.e. a point, parametric and a newly introduced pseudo-continuous approach, using three different datasets. The latter allowed us to evaluate the accuracy and reliability by introducing two different training and testing scenarios. In a first scenario, we used local data for training and testing. In a second scenario, different datasets for training and testing were used. In general, all methods and models showed good to very good performance and PTFs developed with scenario 1 performed better than when using scenario 2. Deriving PTFs from local datasets, if they exist, was demonstrated to be the better option. The difference in methods used to derive the data and differences between soils around the world are the most important reasons for the difference in performance between these two scenarios.

The newly introduced pseudo-continuous PTFs showed superior results in comparison with point and parametric PTFs. The main advantage in using a pseudo-continuous approach lies, however, in its ability to predict water content *at any desired matric potential*, which renders a (pseudo) continuous curve. This contrasts with the traditional point PTFs which only enable to predict water content at *prescribed matric potentials* or to predict the *parameters of a prescribed analytical water retention curve expression*. The pseudo-continuous approach for developing PTFs furthermore greatly facilitates combining several datasets in which water contents have been determined at different matric potentials resulting in a larger useful dataset, in contrast with developing point PTFs. The pseudo-continuous approach finally enables to increase the number of data rows for training the neural network by a factor equal to the number of water retention pairs in the dataset. We therefor recommend the use of pseudo-continuous PTFs to further improve and develop PTFs, in particular when datasets are limited.

Acknowledgments

We are very grateful to Dr. Minasny and Dr. McBratney who gave the registered version of Neuropack software to us. We also deeply acknowledge three anonymous reviewers and Dr. Attila Nemes who's comments and suggestions greatly improved our paper.

References

Assouline, S., Tessier, D., Bruand, A., 1998. A conceptual model of the soil water retention curve. Water Resources Research. 34 (2), 223–231.
Baker, L., Ellison, D., 2008. Optimisation of pedotransfer functions using an artificial neural network ensemble method. Geoderma 144, 212–224.
Blake, G.R., Hartge, K.H., 2002. Bulk density. In: Dane, J.H., Topp, G.C. (Eds.), Methods of Soil Analysis. Part 4. Physical Methods. SSSA Book Ser. 5. American Society of Agronomy, Madison, pp. 363–375.
Breiman, 1996. Bagging predictors. Machine Learning 26, 123–140.
Brooks, R.H., Corey, A.T., 1964. Properties of porous media affecting fluid flow. Journal of Irrigation and Drainage 92, 61–88.
Cornelis, W.M., Ronsyn, J., Van Meirvenne, M., Hartmann, R., 2001. Evaluation of pedotransfer functions for predicting the soil moisture retention curve. Soil Science Society of America Journal 65, 638–648.
Cornelis, W.M., Khlosi, M., Hartmann, R., Van Meirvenne, M., De Vos, B., 2005. Comparison of unimodal analytical expressions for the soil-water retention curve. Soil Science Society of America Journal 69, 1902–1911.
Gee, G.W., Bauder, J.W., 1986. Particle size analysis. In: Klute, A. (Ed.), Methods of soil analysis. Part 1, seconded. Agron., Monogr. 9. ASA and SSSA, Madison, WI, pp. 383–411.
Haghverdi, A., Ghahraman, B., Khoshnud Yazdi, A.A., Arabi, Z., 2010. Estimating of water content in FC and PWP in north and north east of Iran's soil sampels using *k*-nearest neighbor and artificial neural networks. Journal of Water and Soil 24, 804–814 (In Persian).
Jain, S.K., Singh, V.P., van Genuchten, M.Th., 2004. Analysis of soil water retention data using artificial neural networks. Journal of Hydrologic Engineering 9, 415–420.
Jury, W.A., Horton, R., 2004. Soil Physics. John Wiley & Sons, Hoboken, New Jersey.
Khlosi, M., Cornelis, W.M., van Genuchten, M.Th., Douek, A., Gabriels, D., 2008. Performance evaluation of models that describe the soil water retention curve between saturation and oven dryness. Vadose Zone Journal 7, 87–96.
Khoshnood Yazdi, A., Ghahraman, B., 2004. Investigation of relationship between soil texture and scaling parameter to predict soil water content. Journal of Agricultural Engineering Research 20, 17–34 (In Persian).
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 Science Society of America Journal 72, 1243–1247.
Manyame, C., Morgan, C.L., Heilman, J.L., Fatondji, D., Gerard, B., Payne, W.A., 2007. Modeling hydraulic properties of sandy soils of Niger using pedotransfer functions. Geoderma 141, 407–415.
Minasny, B., McBratney, A.B., 2002a. The Neuro-m method for fitting Neural Network parametric pedotransfer functions. Soil Science Society of America Journal 66, 352–361.
Minasny, B., McBratney. A.B., 2002b. Neuropack. Neural Network package for fitting pedotransfer functions. Technical Note. v 1.0. From the Australian Centre for Precision Agriculture. <http://www.usyd.edu.au/su/agric/acpa> Retrieved 16.02.04.
Nemes, A., Schaap, M.G., Wösten, J.H.M., 2003. Functional evaluation of pedotransfer functions derived from different scales of data collection. Soil Science Society of America Journal 67, 1093–1102.
Nemes, A., Rawls, W.J., Pachepsky, Y.A., 2006a. Use of the nonparametric Nearest Neighbor approach to estimate soil hydraulic properties. Soil Science Society of America Journal 70, 327–336.
Nemes, A., Rawls, W.J., Pachepsky, Ya.A., Van Genuchten, M.Th., 2006b. Sensitivity of the Nearest Neighbor approach to estimate soil hydraulic properties. Vadose Zone Journal 5, 1222–1235.

Nemes, A., Quebedeaux, B., Timlin, D.J., 2010. Ensemble approach to provide uncertainty estimates of soil bulk density. Soil Science Society of America Journal 74, 1938–1945.

Or, D., Wraith, J.M., 2002. Soil water content and water potential relationships. In: Warrick, A.W. (Ed.), Soil Physics Companion. CRC Press, Boca Raton, FL, pp. 49–84.

Pachepsky, Y.A., Timlin, D., Várallyay, G., 1996. Artificial neural networks to estimate soil water retention from easily measurable data. Soil Science Society of America Journal 60, 727–733.

Parasuraman, K., Elshorbagy, A., Cheng Si, B., 2006. Estimating saturated hydraulic conductivity in spatially variable fields using Neural Network ensembles. Soil Science Society of America Journal 70, 1851–1859.

Patil, N.G., Rajput, G.S., 2009. Evaluation of water retention functions and computer program "Rosetta" in predicting soil water characteristics of seasonally impounded shrink–swell soils. Journal of Irrigation and Drainage Engineering 135, 286–294.

Schaap, M.G., 2000. Rosetta, Version 1.2. US Salinity Laboratory, USDA, ARS, Riverside, CA. http://www.ussl.ars.usda.gov (verified 28.09.01).

Schaap, M.G., Leij, F.J., 1998a. Using neural networks to predict soil water retention and soil hydraulic conductivity. Soil and Tillage Research 47, 37–42.

Schaap, M.G., Leij, F.J., 1998b. Database-related accuracy and uncertainty of pedotransfer functions. Soil Science 163, 765–779.

Sharma, S.K., Mohanty, B.P., Zhu, J., 2006. Including topography and vegetation attributes for developing pedotransfer functions. Soil Science Society of America Journal 70, 1430–1440.

Tomasella, J., Pachepsky, Y.A., Crestana, S., Rawls, W.J., 2003. Comparison of two techniques to develop pedotransfer functions for water retention. Soil Science Society of America Journal 67, 1085–1092.

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 Science Society of America Journal 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.

USDA, 1982. Procedures for collecting soil samples and methods of analysis for soil survey. Soil Survey Investigations. Report No. 1.

van Genuchten, M.Th., 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>.

van Genuchten, M.Th., 1980. A closed-form equation for predicting the hydraulic conductivity of unsaturated. Soil Science Society of America Journal 43, 892–898.

Vereecken, H., Weynants, M., Javaux, M., Pachepsky, Y., Schaap, M.G., van Genuchten, M.Th., 2010. Using pedotransfer functions to estimate the van Genuchten–Mualem soil hydraulic properties: a review. Vadose Zone Journal 9, 1–26.

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. Journal of Hydrology 251, 123–150.