Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/issn/15375110

Research Paper

Comparison of statistical regression and data-mining techniques in estimating soil water retention of tropical delta soils

Phuong M. Nguyen [a,b,*], Amir Haghverdi [c], Jan de Pue [a,1], Yves-Dady Botula [d], Khoa V. Le [b,e,2], Willem Waegeman [f,3], Wim M. Cornelis [a,1]

[a] Department of Soil Management – Ghent University, Coupure Links 653, 9000, Ghent, Belgium
[b] Department of Soil Science, Can Tho University, 3/2 Street, Ninh Kieu District, Can Tho City, Viet Nam
[c] Department of Environmental Sciences, University of California-Riverside, Riverside, CA, 92521, USA
[d] Department of Natural Resources Management – University of Kinshasa, Democratic Republic of the Congo
[e] Department of Scientific Affairs, Can Tho University, 3/2 Street, Ninh Kieu District, Can Tho City, Viet Nam
[f] Department of Mathematical Modelling, Statistics and Bioinformatics – Ghent University, Coupure Links 653, 9000, Ghent, Belgium

ARTICLE INFO

Article history:
Received 28 January 2016
Received in revised form
14 October 2016
Accepted 14 October 2016
Published online 12 November 2016

Keywords:
Pedotransfer functions
Tropical delta soils
Support vector machines
k-Nearest Neighbours
Multiple linear regression
Artificial neural networks

Although a great number of studies have been devoted to develop and evaluate pedotransfer functions (PTFs), several questions still are to be addressed, particularly pertaining to tropical delta soils which received very little attention. One such question relates to the optimal structural dependency between basic soil properties and soil water retention characteristics (SWRC), which could be formulated by various regression methods. It is hypothesised that data mining techniques provide more accurate SWRC-PTFs than statistical linear regression. However, data-mining techniques are often proven as highly data-demanding techniques. The aim of this study was, therefore, to verify that hypothesis for a limited data set of tropical delta soils by comparing the predictive capabilities of point PTFs and pseudo-continuous (PC) PTFs developed by Multiple Linear Regression (MLR), Artificial Neural Networks (ANN), Support Vector Machine for Regression (SVR), and k-Nearest Neighbours (kNN) methods. The results show that point-PTFs derived from data-mining techniques (i.e. ANN, SVR, kNN) offer accurate and reliable estimation of soil water content at several matric potentials. In case of PC-PTFs, ANN and kNN models outperformed SVR and MLR PTFs in validation phase (RMSE of ANN and kNN PTFs were around 0.05 m^3 m^{-3}, while those of SVR PTFs and MLR PTFs rose up to 0.068 and 0.066 m^3 m^{-3}). Our findings confirm the superiority of data-mining approaches

* *Corresponding author*. Department of Soil Management – Ghent University, Coupure Links 653, 9000 Ghent, Belgium. Fax: +32 9 264 62 47.
E-mail addresses: MinhPhuong.Nguyen@ugent.be, nmphuong@ctu.edu.vn (P.M. Nguyen), amirh@ucr.edu (A. Haghverdi), Jan.DePue@ugent.be (J. de Pue), ydbotula@yahoo.fr (Y.-D. Botula), lvkhoa@ctu.edu.vn (K.V. Le), Willem.Waegeman@ugent.be (W. Waegeman), Wim.Cornelis@ugent.be (W.M. Cornelis).
[1] Fax: +32 9 264 62 47.
[2] Fax: +84 7103 838 474.
[3] Fax: +32 9 264 62 20.
http://dx.doi.org/10.1016/j.biosystemseng.2016.10.013

in modelling the complex system of soil and water, even when a limited data set is available. The non-parametric kNN method, though being constrained in estimating SWRC in pseudo-continuous manner, has great benefits due to its flexibility, simplicity, accuracy and capacity to append new observations.

Nomenclature

θ	Volumetric water content, $m^3\ m^{-3}$
ANN	Artificial neural networks
BD	Bulk density, $Mg\ m^{-3}$
h	Matric head, cm water
kNN	k-Nearest Neighbours
LOO	Leave-one-out cross-validation method
ME	Mean of prediction errors
MLR	Multiple linear regression
OC	Soil organic carbon content, %
PC-PTFs	Pseudo-continuous pedotransfer functions
PTFs	Pedotransfer functions
R^2	Coefficient of determination
RMSE	Root mean squared error
SVM	Support vector machines
SVR	Support vector machines for regression
SWRC	Soil water retention characteristic
VMD	Vietnamese Mekong Delta

1. Introduction

Pedotransfer functions (PTFs) provide an indirect estimation of soil water retention characteristics (SWRC) from readily available or easily measurable basic soil properties, and have therefore emerged as an alternative source of SWRC data for large scale applications of agro-hydrological modelling (Twarakavi, Šimůnek, & Schaap, 2009). Although substantial studies have been devoted to develop and evaluate PTFs, several questions still are to be addressed particularly for paddy soils in the tropical deltas where the interrelationship between soil and water has not been well established (Pachepsky, Rajkai, & Tóth, 2015). One such question relates to the optimal structural dependency between basic soil properties and SWRC, which could be formulated by various regression methods. There are two main categories of regression methods which are widely used for PTF development: statistical regression techniques and data mining or pattern-recognition techniques (Pachepsky & Rawls, 2004; Vereecken et al., 2010).

Regarding the state-of-the-art of SWRC-PTFs, most PTFs derived during the past decades are based on statistical regression methods in which the relationship between the basic soil properties and SWRC are quantified by predefined mathematical equations (e.g., the PTFs of Gupta and Larson (1979); Hodnett and Tomasella (2002); Minasny and Hartemink (2011); Saxton and Rawls (2006)). Statistical regression techniques offer simple, reasonable and well-interpretable models, but are also exposed to several drawbacks: estimation results are heavily biased in case of small sample size; the right form of the regression equation which is usually unknown has to be determined *a priori*; rigorous assumptions about probability distribution of error are not easy to fulfil across the data space; and most importantly, the regression equations need to be redeveloped and republished in case new data become available (Botula, Nemes, Mafuka, Van Ranst, & Cornelis, 2013; Nemes, Rawls, & Pachepsky, 2006; Patil et al., 2013).

Alternative data mining techniques such as Artificial Neural Networks (ANN), k-Nearest Neighbours (kNN), and Support Vector Machines for Regression (SVR) have been introduced as promising tools for PTF development (Botula, Van Ranst, & Cornelis, 2014). Firstly, these techniques have been successfully used for both classification and regression problems in other fields of hydrology. For examples, ANN, SVR and kNN techniques were effectively used to forecast rainfall (Hong & Pai, 2007; Hu, Liu, Liu, & Gao, 2011), water evaporation from soil and free water surfaces (Baydaroğlu & Koçak, 2014), and inflows of water reservoir (Valipour, Banihabib, & Behbahani, 2012; 2013). Due to their high flexibility and accurate predictive performance, data mining techniques have recently gained popularity in unsaturated soil hydrological studies (Botula et al., 2013). These methods have been intensively tested with soils in the temperate (Lamorski, Pachepsky, Slawínski, & Walczak, 2008; Nemes, Rawls, & Pachepsky, 2006; Pachepsky, Timlin, & Varallyay, 1996; Schaap & Leij, 1998; Twarakavi et al., 2009), and arid to semi-arid climates (Bayat et al., 2013; Ebrahimi, Bayat, Neyshaburi, & Zare Abyaneh, 2013; Khlosi, Alhamdoosh, Douaik, Gabriels, & Cornelis, 2016). Only one kNN study was devoted to highly weathered soils in the humid tropics (Botula et al., 2013). All mentioned authors have confirmed the superiority of the used data mining techniques in modelling the interaction of soil and water as a very complex system compared to traditional MLR (Multiple Linear Regression) techniques, although several drawbacks have also been noticed in the same time such as susceptibility to over-fitting, highly data-demanding, and expert knowledge requirement.

In the meantime, Pachepsky, Rawls, and Lin (2013) have noted that the successfulness of certain regression techniques in terms of providing accurate estimations of SWRC is somewhat controlled by type of PTFs, availability of soil variables used in predictive functions, and size and properties of training databases. Indeed, the data used for calibrating/training the PTFs should account for most of the variation that is likely to be encountered in the area where the data are meant to be used, hence large databases of good quality are generally expected for PTF development (Wösten, Pachepsky, & Rawls, 2001). This requirement, however, is hard to be fulfilled in many developing countries in the tropic, where just a few extensive soil-water studies have been done so far. Mayr and Jarvis (1999) also reported that using a small set of

relevant data, if available, is better than using a large and general data set.

Concerning to the PTF's types that are frequently used to estimate SWRC in the literature, three broad groups have been noticed. They are (1) point-based PTFs that predict the water content at specific chosen matric potentials, (2) parameter-based PTFs that estimate the parameters of analytical expressions of the SWRC, e.g. those of Brooks and Corey (1964); Campbell (1974); van Genuchten (1980), and (3) physical–conceptual PTFs that predict soil hydraulic properties based on a soil structural model (Cornelis, Ronsyn, Van Meirvenne, & Hartmann, 2001; Wösten et al., 2001). The latter has not been widely used in practice due to some limitations mentioned in Cornelis et al. (2001).

For modelling purposes, the parameter-based PTFs which offer the prediction of a whole and continuous SWRC are often preferred, since many flux transport models require the complete SWRC as input parameters (Cornelis et al., 2001). However, using statistical validation analysis, many researchers, e.g. Merdun, Çınar, Meral, and Apan (2006); Pachepsky et al. (1996); Tomasella, Pachepsky, Crestana, and Rawls (2003); Vereecken et al. (2010), have noted that point-based PTFs outperformed the parameter-based PTFs in predicting soil water content. The supremacy of point-PTFs could be attributed to the fact that soil water retention at specific matric potentials is controlled by different basic soil properties (Tomasella et al., 2003; Vereecken et al., 2010), and therefore, point PTFs should provide a better combination of these properties and lead to more accurate functions for SWRC estimation. Recently, Haghverdi, Cornelis, and Ghahraman (2012) introduced a new PTF approach, named "pseudo continuous" PTFs (PC-PTFs), which are capable to determine almost continuous SWRCs without using any analytical soil hydraulic expressions. In their PTF, using matric potentials as additional predictor enables to predict the corresponding water content at any desired matric potential. They proved that PC-PTFs derived by the ANN technique were more accurate and reliable than parameter-based PTFs, and slightly better than point-PTFs when a limited data set was available for PTFs' development.

Moreover, due to the very specific nature in terms of physical and hydraulic soil characteristics of tropical delta soils where the main agricultural practice is paddy rice cultivation (Nguyen, Le, & Cornelis, 2014), it is not advisable to utilise PTFs reported so far in the literature to estimate the soil water characteristics in the tropical delta region (Nguyen, Le, Botula, & Cornelis, 2015). Therefore, the main objectives of this study were (1) to develop and validate point PTFs and PC-PTFs using available limited data sets from the tropical Vietnamese Mekong delta (VMD), and (2) to investigate the predictive capability of various regression techniques (i.e., MLR, ANN, SVR and kNN) in estimating SWRC in both point and pseudo-continuous manners. Since the concept of PC-PTFs has only been tested and compared against ANN and SVR techniques for soils in dry and temperate regions (Haghverdi et al., 2012; Haghverdi, Öztürk, & Cornelis, 2014), we believe that testing the predictive power of these approaches together with others as MLR and kNN would be very useful to those that are in need of SWRC data and/or attempt to develop new PTFs for soils in tropical humid region. The novel feature of our study is that it is the first that considers data mining techniques in developing SWRC-PTFs for a variety of soils in a tropical delta where rice paddy cultivation is the dominant agricultural practice. Additionally, the effect of PTF types on the successfulness of regression techniques in PTF development were concomitantly investigated.

2. Materials and method

2.1. Soil data sets

Two data sets collected from the Vietnamese Mekong Delta (VMD) were employed to calibrate and validate the point and PC-PTFs. A first data set of 160 observations which was used to develop the PTFs (so-called training data set) was obtained from previous work (Nguyen et al., 2014). This data set was constructed from a regional field campaign in the period of August 2010 to January 2011 with the aim of covering a wide range of soils that primarily exploited for agricultural production in the tropical lowland delta (mainly paddy rice, but also upland crops such as vegetable, maize, and sugar cane). The locations of the sampling sites of the training data set and the corresponding soil types are shown in Fig. 1.

Records of 160 samples included soil water content measured at eight matric potentials of −1, −3, −6, −10, −20, −33, −100, and −1500 kPa, bulk density (BD), particle-size distribution (sand, silt, clay content), and organic carbon (OC) content, among other properties. Undisturbed soil samples, taken by standard Kopecky rings, were used to determine soil BD, by the core method (Grossman & Reinsch, 2002), and SWRC. The eight points of the SWRC were quantified using sand-boxes and pressure chambers according to the procedures outlined in Cornelis, Khlosi, Hartmann, Van Meirvenne, and De Vos (2005). The disturbed soil samples, which were taken next to the undisturbed sampling sites were used to determine organic carbon content by means of the wet oxidation method (Walkley & Black, 1934), and particle size distribution by sieve-pipette method (Gee & Bauder, 1986).

The performance of the PTFs derived from the training data set was validated by another independent data set, compiled from the study of Le (2003) which was conducted in the same study area. This set includes complete records for 29 samples taken from 10 soil profiles which are representative for major soil groups in VMD (Fig. 1). The physical and chemical soil properties, and SWRC of the test samples were determined by the same methods as mentioned above for the training data set. Originally, the SWRC of test samples were determined at nine matric potentials (i.e., −0.25, −1, −3, −5, −7, −10, −33, −100, −1500 kPa), which are different from the eight matric potentials in the training data set. To address this, we fitted the - with five free parameters - highly flexible van Genuchten (1980) equation (Cornelis et al., 2005) to each of the 29 SWRCs, and determined water contents at those eight matric potentials using the equation with fitted parameters. Tomasella, Hodnett, and Rossato (2000) noted that the PTFs' predictive capacity is related to the similarity between the data sets used in the developing and testing phases of the PTFs, and therefore, using a small but comparable test data set is appropriate for comparing the predictive capacity of PTFs derived by different methods.

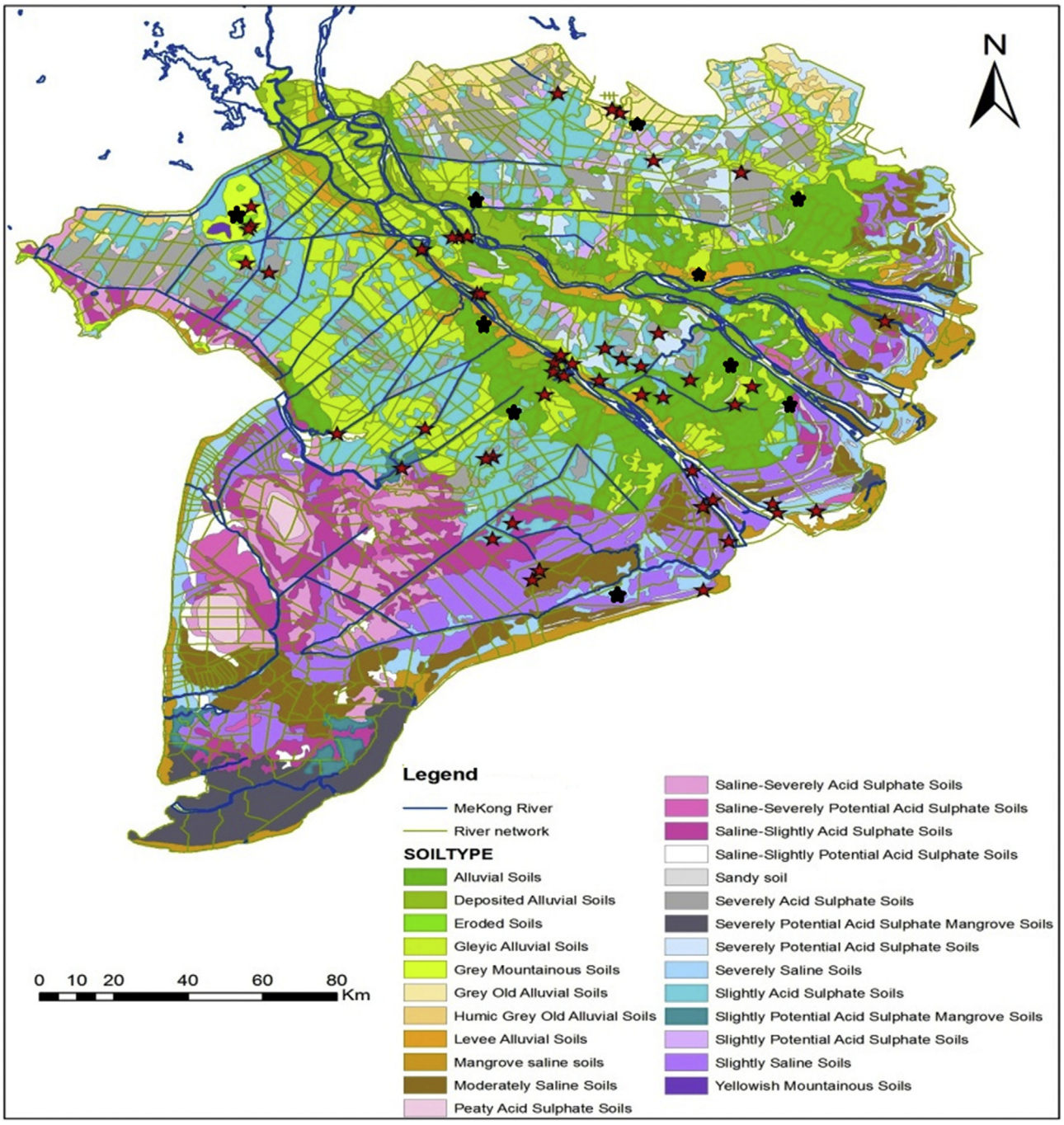


Fig. 1 – The soil map of study area (Vietnamses Mekong Delta) with represented locations of the sampling sites of the training data set (red stars) and the testing data set (black stars). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Descriptive statistics of the basic soil physical, chemical and SWRC of the two aforementioned data sets are summarised in Table 1. The soils in the study area show a wide range of basic soil properties. For example, the range of sand, silt, clay contents, BD and OC content are 0.13–98.6%, 0.00–64.9%, 1.4–76.8%, 0.7–1.9 Mg m^{-3}, and 0.08–12.3%, respectively, for the training data set. The testing data set shows similar ranges. The wide ranges in soil properties are associated with the alluvial nature of the soils. They can be defined as medium- and fine-textured (Fig. 2) with mean percentage of silt and clay of 40.1 and 44.3%, respectively. The soils in the study area are classified as Fluvisols, Gleysols, Luvisols, Acrisols, Arenosols and Plinthosols according to World Rerefernce Base system (IUSS Working Group WRB., 2014), corresponding to Entisols, Inceptisols, and Ultisols of USDA-Soil Taxonomy system (Soil Survey Staff, 1975, p. 436).

2.2. Types of soil water retention Pedotransfer functions

Two types of PTFs were derived to estimate SWRC using various regression techniques, which will be described hereinafter. The first one pertains to point PTFs that estimate soil water content at soil matric potentials of −1, −3, −6, −10, −20, −33, −100, −1500 kPa. The used input variables are the basic soil properties which have been widely used for SWRC estimations (i.e., sand, silt, clay content, BD, and OC content). The second type refers to PC-PTF which uses the logarithm of matric heads as extra input variable, hence supposedly allowing the prediction of soil water content at any desired matric potentials. The structural topologies of the point-based, and pseudo-continuous PTFs are manifested in Fig. 3.

As can be observed in Fig. 3, sand, silt and clay, BD and OC content are the common input predictors of SWRC-PTFs. θ_1 to θ_8 are volumetric water contents which in turn are the outputs of the point PTFs. In case of PC-PTFs, h (in cm water) is matric head and is the extra input variable. θ(h), the output of the PC-PTFs, is volumetric water content at h matric head. Different h values yield different soil water contents. In this study, eight values of h (i.e., 10, 30, 60, 100, 200, 330, 1000, 15,000 cm) corresponding to matric potentials used to derive point PTFs were applied to estimate SWRC using PC-PTFs.

2.3. Methods to derive Pedotransfer functions

2.3.1. Multiple linear regression

The general form of MLR-based PTF is as follows:

$$Y_i = a_1X_1 + a_2X_2 + a_3X_3 + \ldots + a_nX_n$$

where Y_i denotes the response variable (for point-PTFs: Y_i with i = 1, 2, …, 8 is the water content at 8 matric potentials, while in case of PC-PTF, Y is the water content at any corresponding matric head input), X_1, X_2, X_3,… X_n are predictor variables (i.e., sand, silt, clay content, BD and OC content in case of point PTFs, and all aforementioned variables plus logarithm of matric head for the PC-PTF in this study), and a_1, a_2, a_3, … a_n are the values of the regression coefficients obtained from fitting the equation to the training data set. Further detailed explanation about the theory of the MLR method can be found in Kutner, Nachtsheim, Neter, and Li (2005).

The R statistical language (R Core Team, 2014) was used to develop the MLR-based PTFs, in which the potential and significant predictors for SWRC estimation were selected by stepwise regression using Akaike Information Criteria (AIC) as selection criterion.

2.3.2. Artificial neural networks

The point and pseudo-continuous ANN-PTFs were established in Matlab R2014a environment (MathWorks, Natick, MA, USA). A three-layer feed forward back propagation ANN model was selected. The activation functions were sigmoid tangent hyperbolic and linear in hidden and output layers, respectively. The Levenberg–Marquardt algorithm (Demuth & Beale, 2000) was implemented for training. The number of neurons in the hidden layer was changed from 1 to 20. The input and outputs of the ANN-PTFs were identical to those of MLR-PTFs. The bootstrap method (Efron & Tibshirani, 1993) was applied on training data to create 50 statistically similar subsets of the same size through a sampling with replacement technique. The subsets, comprising about 63% of the parent data (Schaap, Leij, & van Genuchten, 2001), were used to train the PTFs. The idle samples (i.e. 37%) formed the cross-validation set. The training was stopped whenever the error increased on the cross-validation set. The outputs of the 50 bootstraps were averaged and reported as the predictions from the PTFs.

2.3.3. Support vector machines for regression

The SVM algorithm was implemented in the R statistical language (R Core Team, 2014) to derive point and PC-PTFs. The most commonly used kernel, i.e. radial basis function kernel, which has been applied in the works of Haghverdi et al. (2014); Lamorski et al. (2008); Twarakavi et al. (2009) was selected to build our SVR-models. The optimal hyper-parameters of the SVR models were estimated using a thorough grid-based search approach (Hastie, Tibshirani, & Friedman, 2009) in which the parameter C was changed from 0.001 to 1 in increment of 0.1, ε was varied from 0 to 1 in increment of 0.05, while γ was adjusted from 0.01 to 1 with a mesh increment of

Table 1 – Descriptive statistics of soil properties in the training (N = 160) and test (N = 29) data sets.

Soil properties	Training data set				Test data set			
	Min	Max	Mean	Std	Min	Max	Mean	Std
Organic carbon (%)	0.08	12.3	2.37	2.41	0.03	7.75	1.17	1.53
Bulk density (Mg m^{-3})	0.7	1.9	1.25	0.24	0.83	1.81	1.31	0.26
Sand content (%)	0.1	99	15.6	26.8	1	80	12.1	23.1
Silt content (%)	0.0	65	40.1	13.8	5	56	38.8	12.3
Clay content (%)	1	77	44.3	19.0	3	67	47.5	17.9
θ (m^3 m^{-3}) at:								
−1 kPa	0.24	0.74	0.50	0.10	0.31	0.66	0.52	0.08
−3 kPa	0.17	0.73	0.49	0.10	0.28	0.65	0.51	0.09
−6 kPa	0.12	0.72	0.47	0.12	0.26	0.62	0.50	0.09
−10 kPa	0.06	0.71	0.45	0.12	0.22	0.59	0.49	0.10
−20 kPa	0.03	0.70	0.41	0.12	0.17	0.55	0.46	0.10
−33 kPa	0.03	0.67	0.37	0.12	0.14	0.51	0.42	0.10
−100 kPa	0.03	0.58	0.32	0.11	0.08	0.43	0.35	0.09
−1500 kPa	0.02	0.43	0.24	0.09	0.04	0.25	0.21	0.06

Min, Max, Std are the minimum, the maximum and the standard deviation of soil variables.

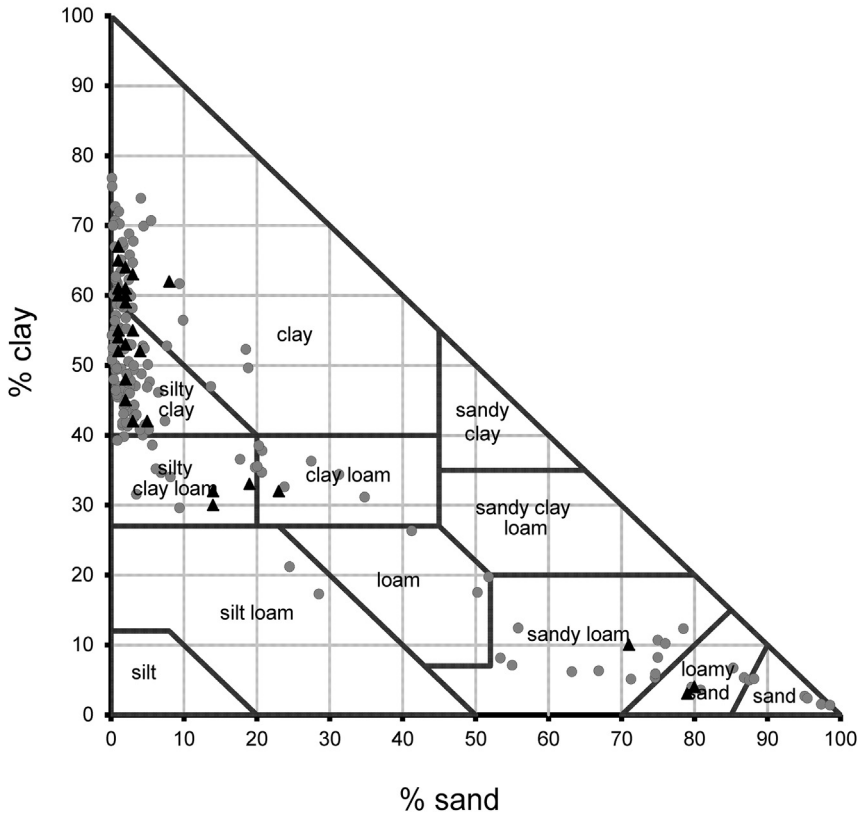


Fig. 2 – Variation of clay and sand content in the training data set (grey circles) and the test data set (black triangles) within the USDA textural triangle.

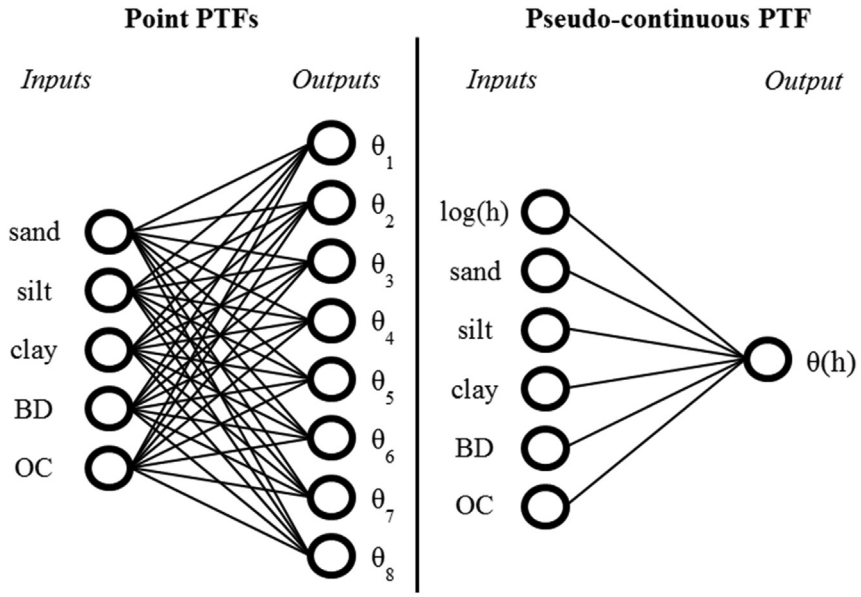


Fig. 3 – Structural topologies of the point and pseudo-continuous PTFs used in this study (modified from Haghverdi et al., 2012). The input (predictor) variables sand, silt and clay are sand, silt and clay percentages, BD is bulk density in Mg m^{-3}, OC is organic carbon percentage, log(h) is the common logarithm of the absolute value of matric head h in cm water. The output (response) variables θ_1, θ_2, … θ_8 are volumetric water contents in m^3 m^{-3} at eight different matric heads (see Table 1) in case of the point PTFs, whereas θ(h) is the volumetric water content at any preselected matric head h in case of the pseudo-continuous PTF; in this study, preselected h values corresponded to those of the point PTF.

0.1. The range of the meta-parameters was obtained from the preliminary grid searches with large mesh of increment. A two-round grid search was implemented due to the high computational cost of the SVR optimisation process for three parameters simultaneously. Ten-fold cross validation technique was used in the grid-based search; and the set of C, γ, ε corresponding to the best cross-validation accuracy was picked and used to calibrate SVR models.

Detailed information about the methodology of this technique can be found in many previous works, e.g., Lamorski et al. (2008); Smola and Schölkopf (2004); Twarakavi et al. (2009).

2.3.4. *k-Nearest Neighbours*

The basic idea of the kNN technique, named similarity-based technique by Nemes, Rawls, and Pachepsky (2006), is finding the k Nearest Neighbours from a reference dataset for each soil in the test dataset in terms of selected input attributes. For kNN estimation, two design parameters were defined and used for the estimation procedure, namely the k and p terms. The k term refers to the number of similar soils to be selected from the reference data set to estimate the output attributes for each target soil, while the p term determines the weight–distance relationship that determines the contribution of each of the k reference samples to the estimation of the output attribute, depending on their degree of similarity to the target soil. Working with extensive soil databases in temperate regions, Nemes, Rawls, and Pachepsky (2006) derived regression equations relating the k and p terms with training data set size. Botula et al. (2013) tested this relation for soils in the humid tropics and obtained similar results. Therefore, in this study, we use the proposed formula of Nemes, Rawls, and Pachepsky (2006) for determining the designed parameters k and p.

More methodological and calculation details on the whole procedure can be found in the works of Botula et al. (2013); Nemes, Rawls, and Pachepsky (2006); Nemes, Rawls, Pachepsky, and van Genuchten (2006). The kNN algorithm used in this study was adapted from the variants developed by Nemes, Rawls, and Pachepsky (2006) and Botula et al. (2013). The implementation of the kNN algorithm was done in the Matlab R2014a environment.

2.4. *Evaluation criteria*

The evaluation of PTF performance, in terms of accuracy and reliability, is commonly made and reported through the comparison of PTF estimated and observed values. As it was clearly stated by Wösten et al. (2001), accuracy of a PTF is defined as the correspondence between observed and predicted data in the training data set, whereas reliability of PTFs is assessed in terms of correspondence between observed and predicted data in other independent data sets. In the present study, three statistical indices, i.e. (1) mean of prediction error (ME), a measure of the prediction bias which indicates the over- or under-estimations of a specific model, (2) the root mean square of the prediction error (RMSE), a measure of the overall prediction error, and (3) the coefficient of determination (R^2) which indicates the amount of variation in the data explained by the regression model, were selected to assess the predictive ability of the derived PTFs in both calibration and validation phases. These statistical indices were calculated using the following equations:

$$ME = \frac{1}{N}\sum_{i=1}^{N}(E_i - O_i)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(E_i - O_i)^2}$$

$$R^2 = \frac{\sum_{i=1}^{N}\left[(E_i - \overline{E})\cdot(O_i - \overline{O})\right]^2}{\sum_{i=1}^{N}(E_i - \overline{E})^2 \cdot \sum_{i=1}^{N}(O_i - \overline{O})^2}$$

where E_i is the ith estimated value, O_i is the ith observed value, $\overline{E}$ is the mean of estimated values, $\overline{O}$ is the mean of observed values, and N is the number of observations.

For the MLR and SVR approaches, all soil samples in the training data set were used to evaluate the accuracy of the derived PTFs (N is equal to 160 and 1280 observations for point and PC-PTFs, respectively). For the kNN algorithm, the accuracy of the models was evaluated based on the leave-one-out (LOO) procedure as it is an instance-based regression method which needs separate test data for evaluation. As the name suggests, during the evaluation procedure, one sample was left out and the remaining samples were used as the training data to derive the estimation for the leave-out sample (Mucherino, Papajorgji, & Pardalos, 2009). Therefore, $N = 160 - 1 = 159$ observations for point estimation. In case of PC_{kNN}-PTF, eight corresponding water retention points of the leave-out soil were removed, hence $N = 1280 - 8 = 1272$ observations for pseudo-continuous estimation. The reliability of point PTFs derived by these regression methods was assessed by using the test data set of $N = 29$ samples, corresponding to $N = 29 \times 8 = 232$ samples for PC-PTFs.

3. Results

3.1. *Exploratory data analysis*

In order to find interrelations of soil properties in the training data set, an exploratory data analysis was conducted. The Pearson correlation matrix (Table 2) displayed significant correlation between soil moisture retained at different matric potentials and particle size distribution (i.e., sand, silt, clay content), BD, and OC content. Clay, silt and OC content were positively correlated with soil water content, whereas BD and sand content were negatively correlated. The correlation strength between OC or BD, and soil moisture increased with increasing soil matric potentials (i.e., less negative), whereas those of clay or sand content rose with decreasing matric potentials. These observations were expected because soil structure which determines water content stored in soils at high potentials is more related to OC content and BD (Botula et al., 2013; Pachepsky, Rawls, & Lin, 2006), whereas soil water content at the dry end of the SWRC is primarily determined by adsorption forces in the soil matrix (mainly determined by soil texture, particularly clay content) (Manrique, Jones, & Dyke, 1991). Additionally, there

Table 2 – Pearson's correlation coefficients between soil properties in the training data set (N = 160).

Soil properties	clay	silt	sand	BD	OC	logOC	$\theta_{-1\ kPa}$	$\theta_{-3\ kPa}$	$\theta_{-6\ kPa}$	$\theta_{-10\ kPa}$	$\theta_{-20\ kPa}$	$\theta_{-33\ kPa}$	$\theta_{-100\ kPa}$	$\theta_{-1500\ kPa}$
clay	1													
silt	0.33**	1												
sand	−0.87**	−0.74**	1											
BD	−0.47**	−0.23**	0.45**	1										
OC	0.35**	0.07	−0.28**	−0.75**	1									
logOC	0.61**	0.33**	−0.6**	−0.80**	1.00**	1								
$\theta_{-1\ kPa}$	0.71**	0.27**	−0.64**	−0.72**	0.65**	0.75**	1							
$\theta_{-3\ kPa}$	0.74**	0.31**	−0.68**	−0.72**	0.64**	0.77**	0.99**	1						
$\theta_{-6\ kPa}$	0.80**	0.41**	−0.77**	−0.68**	0.62**	0.80**	0.94**	0.97**	1					
$\theta_{-10\ kPa}$	0.82**	0.46**	−0.82**	−0.65**	0.59**	0.80**	0.90**	0.93**	0.99**	1				
$\theta_{-20\ kPa}$	0.84**	0.51**	−0.85**	−0.63**	0.54**	0.76**	0.84**	0.88**	0.95**	0.98**	1			
$\theta_{-33\ kPa}$	0.82**	0.53**	−0.86**	−0.61**	0.49**	0.71**	0.79**	0.82**	0.90**	0.94**	0.99**	1		
$\theta_{-100\ kPa}$	0.80**	0.56**	−0.85**	−0.56**	0.40**	0.65**	0.71**	0.75**	0.82**	0.87**	0.94**	0.98**	1	
$\theta_{-1500\ kPa}$	0.84**	0.54**	−0.87**	−0.49**	0.33**	0.60**	0.7**	0.74**	0.81**	0.85**	0.90**	0.93**	0.95**	1

** shows significant correlation at 0.01 significance level.

Table 3 – Mean prediction error (ME) and coefficient of determination (R^2) of point PTFs derived based on various regression techniques, i.e., multiple linear regression (MLR), artificial neural networks (ANN), support vector machines for regression (SVR), and k-Nearest Neighbours (kNN).

Matric potentials (kPa)	Training phase				Testing phase			
	MLR	ANN	SVR	kNN	MLR	ANN	SVR	kNN
	ME ($m^3\ m^{-3}$)							
−1	$-8.7*10^{-17}$	0.0013	−0.0004	−0.0002	−0.040	−0.032	−0.034	−0.028
−3	$-1.8*10^{-16}$	0.0014	0.0006	−0.0007	−0.046	−0.036	−0.029	−0.031
−6	$-1.1*10^{-16}$	0.0011	−0.0013	−0.0016	−0.066	−0.043	−0.035	−0.037
−10	$-6.3*10^{-17}$	0.00002	0.0029	−0.0018	−0.075	−0.050	−0.043	−0.042
−20	$-7.6*10^{-17}$	0.0002	−0.0009	−0.0024	−0.071	−0.056	−0.054	−0.049
−34	$-3*10^{-17}$	0.0014	−0.003	−0.0026	−0.061	−0.054	−0.056	−0.049
−100	$-8*10^{-17}$	0.0006	−0.006	−0.0021	−0.024	−0.034	−0.040	−0.028
−1500	$7.4*10^{-16}$	0.0001	−0.0004	−0.0005	0.043	0.033	0.035	0.041
Average	**$1.4*10^{-17}$**	**0.0008**	**−0.001**	**−0.001**	**−0.043**	**−0.034**	**−0.032**	**−0.028**
	R^2							
−1	0.71	0.78	0.79	0.68	0.84	0.83	0.88	0.89
−3	0.75	0.81	0.78	0.71	0.84	0.83	0.88	0.90
−6	0.80	0.84	0.79	0.76	0.79	0.83	0.92	0.90
−10	0.83	0.87	0.82	0.79	0.78	0.83	0.87	0.90
−20	0.84	0.88	0.83	0.79	0.84	0.81	0.88	0.89
−34	0.81	0.86	0.81	0.76	0.87	0.82	0.88	0.89
−100	0.77	0.83	0.82	0.73	0.91	0.83	0.91	0.88
−1500	0.79	0.83	0.81	0.76	0.85	0.79	0.82	0.88
Average	**0.79**	**0.84**	**0.81**	**0.75**	**0.84**	**0.82**	**0.88**	**0.89**

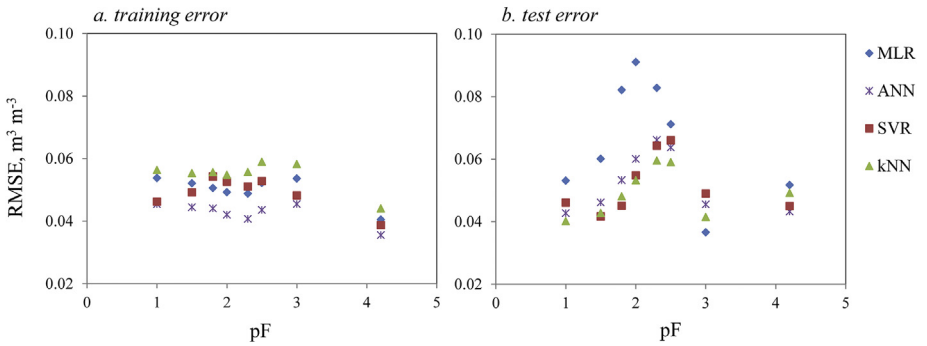


Fig. 4 – Variation of root mean square error (RMSE) as a function of pF (pF is log(h) where h is matric head expressed in cm water) in training phase (a) and testing phase (b) of point PTFs derived by multiple linear regression (MLR), artificial neural networks (ANN), support vector machines for regression (SVR), and k-Nearest Neighbours (kNN) methods.

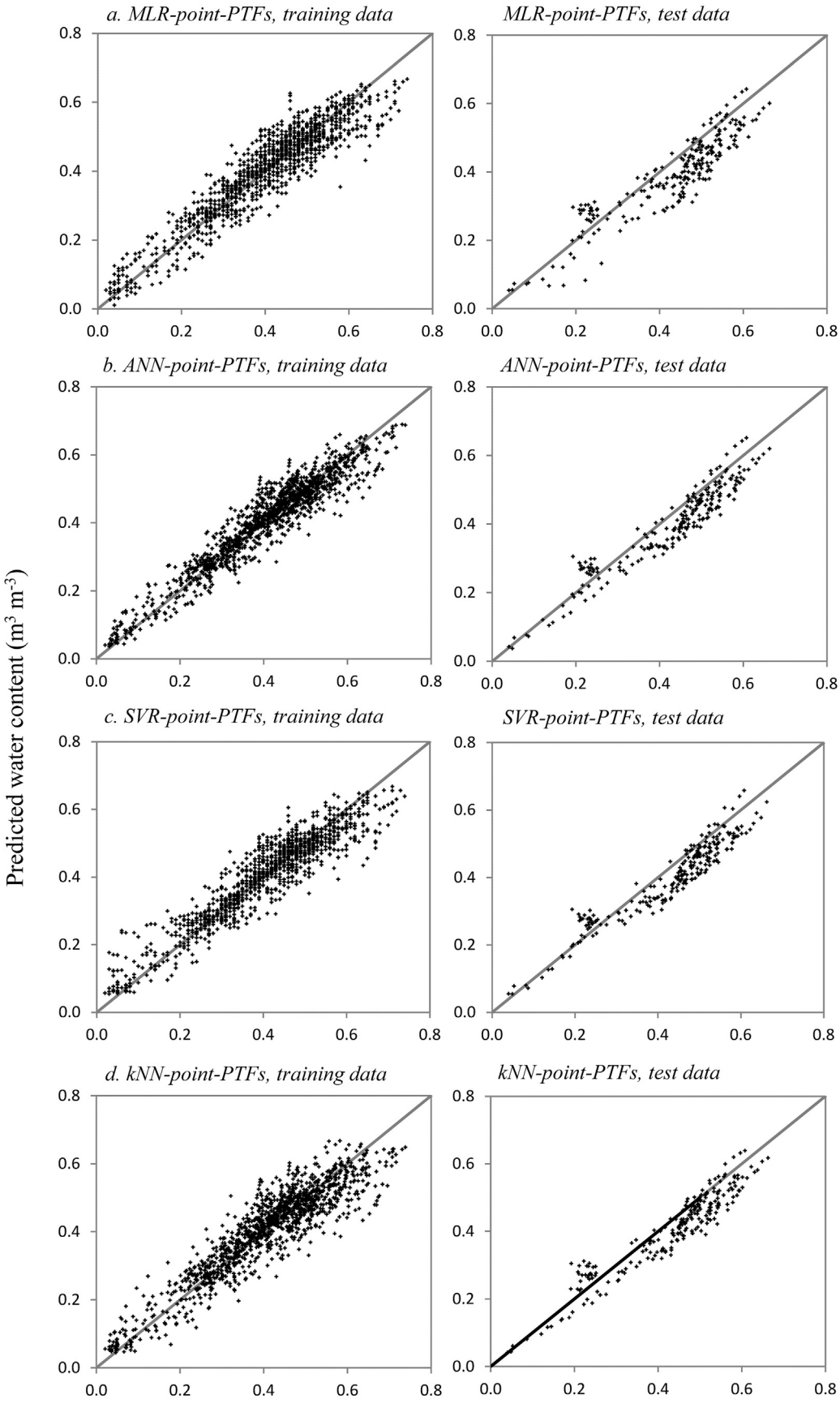

a. MLR-point-PTFs, training data
MLR-point-PTFs, test data
b. ANN-point-PTFs, training data
ANN-point-PTFs, test data
c. SVR-point-PTFs, training data
SVR-point-PTFs, test data
d. kNN-point-PTFs, training data
kNN-point-PTFs, test data
Predicted water content (m3 m-3)
Observed water content (m3 m-3)
0.0
0.2
0.4
0.6
0.8

Table 4 – Performance evaluation using different statistical indices as RMSE, ME and R^2 of point PTFs and PC-PTFs (bold numbers) developed by multiple linear regression (MLR), artificial neural networks (ANN), support vector machines for regression (SVR) and k-Nearest Neighbours (kNN) methods.

Regression methods		RMSE		ME		R^2	
		Point PTFs	PC-PTFs	Point PTFs	PC-PTFs	Point PTFs	PC-PTFs
Accuracy	MLR	0.050	**0.056**	$1.4*10^{-17}$	$-3*10^{-18}$	0.79	**0.84**
	ANN	0.043	**0.044**	0.0008	**−0.0007**	0.84	**0.90**
	SVR	0.049	**0.036**	−0.001	**−0.0009**	0.81	**0.93**
	kNN	0.055	**0.056**	−0.001	**0.0003**	0.75	**0.84**
Reliability	MLR	0.068	**0.066**	−0.043	**−0.043**	0.84	**0.85**
	ANN	0.053	**0.052**	−0.034	**−0.035**	0.82	**0.90**
	SVR	0.052	**0.068**	−0.032	**−0.044**	0.88	**0.84**
	kNN	0.049	**0.050**	−0.028	**−0.027**	0.89	**0.90**

is strong correlation between OC content and soil bulk density ($r = -0.75$), confirming that soils with higher organic matter will concomitantly have lower bulk density. Similarly, Botula et al. (2013) found a significant though less pronounced ($r = -0.382$) correlation between OC and BD of highly weathered soils in the humid tropics, whereas for temperate forest soils De Vos, Van Meirvenne, Quataert, Deckers, and Muys (2005) showed that loss-on-ignition alone was able to explain 57% of the variation in bulk density (corresponding with $r = -0.75$). On the other hand, significant but weak correlations were observed between clay, sand contents and OC. This is most probably because the training data were collected in the region where paddy rice is the main agricultural practice. Rice soils, coincident with fine-textured soils, often have high OC accumulation in the surface due to long-lasting submerged condition of paddy-rice cultivation (Linh et al., 2015).

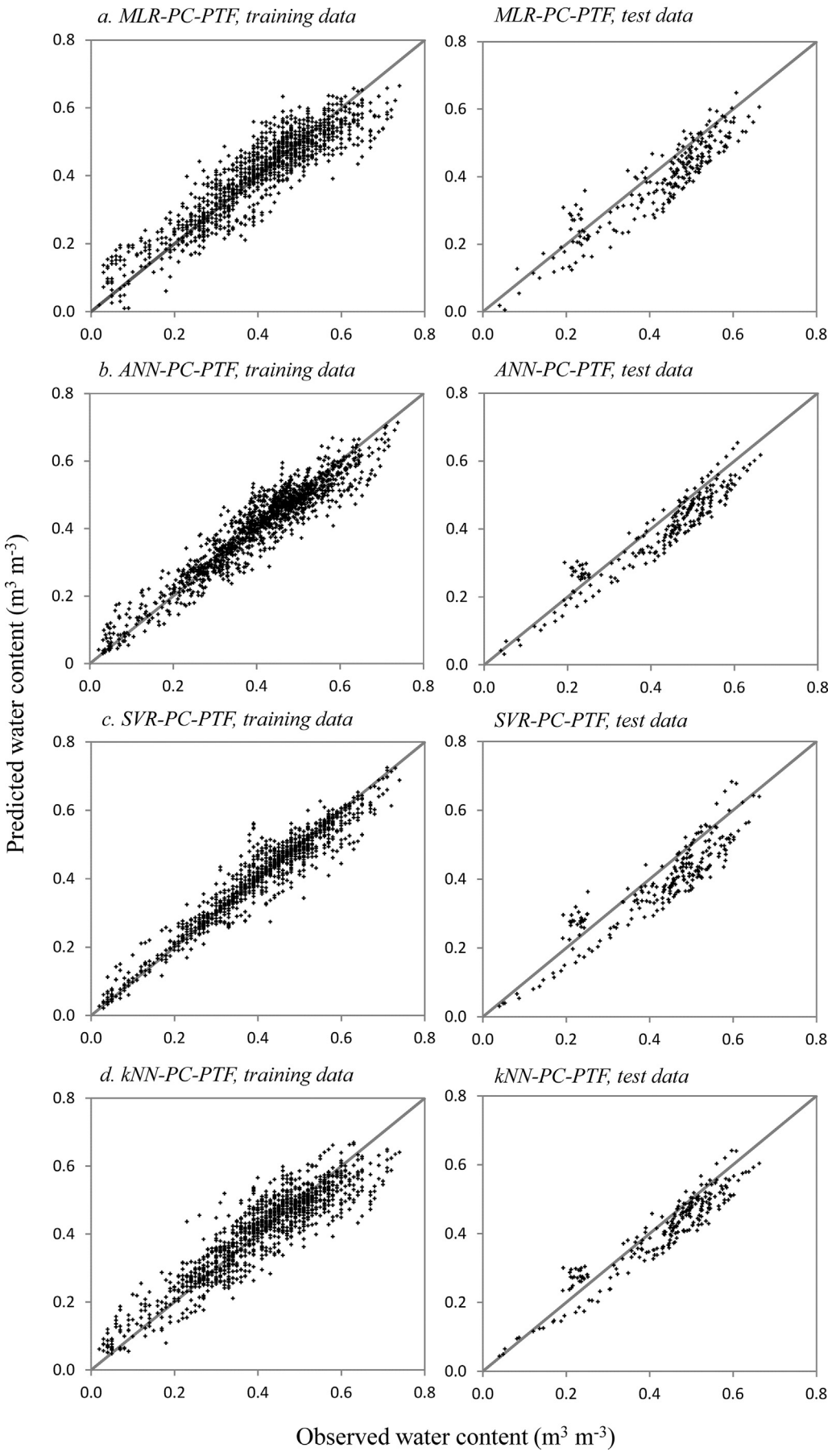
The exploratory data analysis also exposed an exponential relationship between total OC content and soil moisture content retained at different pressure heads in both data sets. In order to properly deriving optimal PTFs based on the linear regression technique, a log-transformation was applied to resolve the non-linearity problem of total OC content before conducting MLR analysis. Data mining techniques, on the other hand, can theoretically handle highly non-linear data, and therefore original soil variables in the training data set were used to train ANN, SVR and kNN models.

3.2. *Point Pedotransfer functions performance*

The R^2, ME, and RMSE of point PTFs derived by different methods are summarised in Table 3 and Fig. 4. All of the four prediction methods displayed a comparable performance in the training phase, in which the ANN-PTFs performed best. The relative high value of coefficient of determination (average R^2 equals 0.79, 0.84, 0.81, and 0.75 for MLR-, ANN-, SVR- and kNN-PTFs, respectively) indicates that a large proportion of the SWRC variability of training samples can be explained by these empirical models using basic soil data (i.e., soil texture, BD and OC content). Similarly, the log(h) dependent analysis of RMSE (left panel of Fig. 4) displays satisfactory accurate estimations of soil water content at different pressure heads with RMSE ranges of 0.040–0.054 $m^3\ m^{-3}$ for MLR-PTFs, 0.036–0.045 $m^3\ m^{-3}$ for ANN-PTFs, 0.039–0.053 $m^3\ m^{-3}$ for SVR-PTFs, and 0.044–0.059 $m^3\ m^{-3}$ for kNN-PTFs. The scatter plots of observed *vs.* PTF predicted soil water content of different models (left panel of Fig. 5), together with approximately-closed zero values of ME in the training phase (Table 3), once again strengthen the appropriateness of these methods in describing SWRC of tropical lowland delta soils. Most of the points are closely scattered around the reference 1:1 lines and do not exhibit much bias.

Further detailed assessment of all error measures (i.e., ME, RMSE and R^2) of these methods (Table 3 and Fig. 4), however, expose a marginally worse performance of kNN-PTFs compared to ANN-, SVR- and MLR-PTFs. The worse performance of kNN models in the training phase as compared with MLR models can be explained by the nature of these regression techniques. In fact, the MLR model, fitted by least square method, makes a huge assumption about the structure of the empirical relationship; hence it yields a stable but possibly inaccurate prediction (i.e., low variance and potentially high bias). Whereas, the kNN model makes very mild structural assumptions; therefore, its predictions are often accurate but can be unstable (i.e., high variance and low bias) (Hastie et al., 2009). For kNN fits, the error on the training data is approximately an increasing function of the *k* parameter; therefore, an independent test set would give us a more satisfactory means for comparing the different methods. Obviously, kNN as well as other pattern recognition techniques like ANN and SVR were more reliable than MLR models in predicting soil water retention of independent test samples (Table 3 and right panel of Figs. 4 and 5). The average values of RMSE, ME and R^2 for the validation data set equal 0.049 $m^3\ m^{-3}$, −0.028 $m^3\ m^{-3}$, 0.89 for kNN-PTFs; 0.052 $m^3\ m^{-3}$, −0.032 $m^3\ m^{-3}$, 0.88 for SVR-PTFs; and 0.053 $m^3\ m^{-3}$, −0.034 $m^3\ m^{-3}$, 0.82 for ANN-PTFs,

Fig. 5 – Scatter plots of observed *vs.* predicted water content ($m^3\ m^{-3}$) of point PTFs for training and test data sets. MLR-point-PTFs (a) are the point PTFs derived by multiple linear regression, ANN-point-PTFs (b) are the point PTFs derived by artificial neural networks, SVR-point-PTFs (c) are the point PTFs derived by support vector machines for regression, kNN-point-PTFs (d) are the point PTFs derived by k-Nearest Neighbours method.


a. MLR-PC-PTF, training data
MLR-PC-PTF, test data
b. ANN-PC-PTF, training data
ANN-PC-PTF, test data
c. SVR-PC-PTF, training data
SVR-PC-PTF, test data
d. kNN-PC-PTF, training data
kNN-PC-PTF, test data
Predicted water content (m3 m-3)
Observed water content (m3 m-3)
0.0
0.2
0.4
0.6
0.8

respectively. These results asserted that the reliability of point PTFs derived by the data mining or pattern-recognition approaches are much better than that of MLR-PTFs (RMSE = 0.068 $m^3 m^{-3}$, ME = −0.043 $m^3 m^{-3}$, R^2 = 0.84).

The variation of RMSE in relation to soil matric potentials (right panel of Fig. 4) of all four methods exposed a similar trend (i.e., RMSE was relatively low at the wet part of the curve, increased toward the intermediate region, but decreased again at the dry part) with the extreme climaxes observed with MLR-PTFs. Comparable patterns of error variation were also reported by many researchers (Botula et al., 2013; Haghverdi, Leib, & Cornelis, 2015; Vereecken et al., 2010). These authors asserted that the accuracy of PTFs in dependence on the matric potentials are affected by the PTF type, the data characteristic, and the input attributes combination.

3.3. *Pseudo-continuous Pedotransfer functions*

The predictive performance in terms of ME, RMSE and R^2 of PC-PTFs derived by different regression methods is summarised in Table 4. Average statistical indexes of point PTFs derived by corresponding techniques were concomitantly presented for the comparison with PC-PTFs performance.

The evaluation of the accuracy and reliability show that the SVR and MLR techniques are probably not a proper choice of tools for the development of PC-PTFs, at least within the context of this study.

Using the MLR technique, the accuracy of derived PC-PTF in terms of overall prediction error is worse (RMSE = 0.056 $m^3 m^{-3}$) than that of point PTFs (RMSE = 0.05 $m^3 m^{-3}$). The points of the scatter plot of PC_{MLR}-PTFs (Fig. 6) are more dispersed with a sign of underestimation in the wet moisture range and overestimation in the dry moisture range compared to those of point PTFs derived by the same method (Fig. 5). The reliability of PC_{MLR}-PTF in terms of validation error (RMSE = 0.066 $m^3 m^{-3}$) is comparable with the average value of $point_{MLR}$ PTFs (RMSE = 0.068 $m^3 m^{-3}$).

Regarding SVR, the results show that the PC_{SVR}-PTF derived based on the optimal SVR meta-parameters obtained from the 10-fold cross validation process (i.e., C = 1, γ = 0.6, ε = 0.05) offers a good fit to the training data set with R^2 = 0.93, ME = −0.0009 $m^3 m^{-3}$, and RMSE = 0.036 $m^3 m^{-3}$. The scatter plot of PC_{SVR}-PTF displays a substantial agreement between observed and PC-PTF predicted water content (Fig. 6). However, the validation result of the derived PC_{SVR}-PTF exposes a poor generalisation performance to the test samples. The validation error is rather high (RMSE = 0.068 $m^3 m^{-3}$) and in the same order of magnitude with the one yielded by PC_{MLR}-PTF.

The PC-PTF approach was recommended for limited available training data when developed by highly data-demanding techniques (Haghverdi et al., 2012). The suitability of the ANN method in developing PC-PTFs, which was reported in the studies of Haghverdi et al. (2012, 2014) for soils in dry regions, was confirmed in this study for soils in the tropical humid delta in both training (R^2 = 0.90, ME = −0.0007 $m^3 m^{-3}$, and RMSE = 0.044 $m^3 m^{-3}$) and testing phase (R^2 = 0.90, ME = −0.035 $m^3 m^{-3}$, and RMSE = 0.052 $m^3 m^{-3}$).

As regards the kNN method, which was actually applied in this study for the first time to develop PC-PTF, the accuracy of PC_{kNN}-PTF (RMSE = 0.056 $m^3 m^{-3}$) was comparable to that of $point_{kNN}$ PTFs (RMSE = 0.055 $m^3 m^{-3}$). A similar agreement in terms of R^2 and graphical correspondence between measured and predicted water content was also observed (Table 4 and Fig. 6). The reliability of such PC_{kNN}-PTF on the test samples was as good as the $point_{kNN}$ PTFs (RMSE = 0.05 $m^3 m^{-3}$, and R^2 = 0.9).

Regardless of PTF types and regression methods, all derived PTFs underestimate the soil water content of the test samples (ME < 0; right panel in Figs. 5 and 6). Although soils in the test data set came from the same population as the training data, the bias of the estimation might probably be due to temporal variation of soil hydraulic characteristics as a result of changes in land use types and soil management (Or & Wraith, 2002). Nonetheless, the predicted SWRC from outperforming PTFs (i.e., PTFs derived by SVR for point estimation, ANN and kNN for both point and pseudo-continuous estimation) are of acceptable accuracy for indirect estimation approaches, as these RMSE values (Table 4) are in the typical RMSE ranges reported by Vereecken et al. (2010).

4. Discussion

The generalisation strength of data mining techniques (i.e., ANN, kNN and SVM) reported in previous studies, e.g., Botula et al. (2013); Haghverdi et al. (2012); Patil et al. (2013); Twarakavi et al. (2009) was confirmed in this study for point estimation of SWRC of tropical delta soils. Indeed, the pattern recognition techniques of SVR, kNN and ANN do not appear to rely on any stringent assumption about the underlying data and can adapt to any situation, hence providing flexible and reliable estimation (Hastie et al., 2009).

Good generalisation performance of the point PTFs derived by the SVR technique might probably result from the implementation of a structural risk minimisation in the optimisation algorithm (i.e. minimising a test error by controlling two contradictory factors: a risk function from empirical data and a capacity for the set of real-valued functions). This aspect leads to a better generalisation capacity of SVR models for new samples as compared to the statistical linear regression models which employed only empirical risk minimisation (i.e. minimising a training error by maximising the goodness-of-fit to the training data). Also, the satisfactory generalisation

Fig. 6 – Scatter plots of observed vs. predicted water content ($m^3 m^{-3}$) of PC-PTF for training and test data sets. MLR-PC-PTF (a) is the PC-PTF derived by multiple linear regression, ANN-PC-PTF (b) is the PC-PTF derived by artificial neural networks, SVR-PC-PTF (c) is the PC-PTF derived by support vector machines for regression, and kNN-PC-PTF (d) is the PC-PTF derived by k-Nearest Neighbours method.

performance of kNN-PTFs in estimating SWRC of independent test samples would possibly be explained by the similarity-based nature of the kNN model together with the likeliness of SWRC of soils in both training and testing data sets (i.e. they were collected from the same study area – VMD). Indeed, Perkins and Nimmo (2009) and Botula et al. (2013) have stressed that the predictive capability of PTFs derived by pattern recognition techniques depends on the quality and the representability of the training data to soils for which one needs to predict SWRC. The well-defined ability of the ANN technique to mimic the inputs–outputs relationship of complex soil water systems (Pachepsky & Schaap, 2004) might probably explain the adequate performance of ANN-PTFs in both training and testing phases of point and pseudo-continuous estimation. Inversely, the MLR models were constructed based on rigorous structural assumptions of the relationship between SWRC and other soil variables. Hence, the regression equations yield stable but possibly inaccurate estimation (Hastie et al., 2009). It is manifested by poor results of $point_{MLR}$ PTFs with test data in this study.

It is important, however, not to overemphasise the generalisation performance of data mining techniques. The prediction capacity of PC-PTFs derived by SVR method for new test samples is comparable with that of statistical linear regression models. The poor generalisation performance of PC_{SVR}-PTF in this study opposes to the strengths of SVM algorithms (i.e., promising generalisation performance and capacity to handle with non-linear data), which have been reclaimed by other researchers (Lamorski et al., 2008; Twarakavi et al., 2009) for other PTF types (i.e., point-PTFs and parameter-based PTFs). The counter result in the present study is supported by the study of Haghverdi et al. (2014). They noted that for the application of the SVM technique, using only the statistical mean square error (MSE) index to select the optimal model is insufficient for PC-PTF type, because the models that show satisfactory values of mean square error in the training phase are the ones displaying a linear relationship between soil-water content and the logarithm of soil matric head. The reliability of PC_{SVR}-PTF is therefore similar to that of PC_{MLR}-PTF.

Although in contrast with MLR and SVR models, both the point and pseudo-continuous kNN models provide better estimation of SWRC when using independent target samples, it is important to note the limitation of the kNN method in developing PC-PTF. Unlike other parametric regression techniques (i.e., MLR, ANN, SVM) which define relationships of soil properties under mathematical functions, the kNN is a non-parametric regression technique which is limited in its capacity to provide a continuous prediction of SWRC. Moreover, as the concept of PC-PTFs in combination with the kNN technique was applied for the first time in this study, we would like to clarify the difference of this PC_{kNN}-PTF from that of Nemes, Rawls, and Pachepsky (2006). Since the pseudo-continuous topology considers matric potential as extra input variable, the PC-PTF could consider multiple water retention points of a particular soil in the reference/training data which has basic soil properties very similar to the target soils as nearest neighbours. Such selection opposes to the classical kNN-PTF for point estimation (i.e., selecting multiple soils). Therefore, it should actually be considered as point PTF with matric head as additional input variable for the estimation of several points of SWRC. A continuous SWRC can then be obtained by fitting analytical equations to multiple predicted water retention points. Recently, Haghverdi et al. (2015) have introduced kNN-VG-PTFs in which a non-parametric kNN technique was applied in combination with the van Genuchten model. By this way, any points of SWRC of the target sample could be estimated based on the VG parameters of nearest samples withdrawn from the reference database. Such PTFs showed reasonable accuracy and reliability in comparison with other well-known parameter based PTFs. However, as we already presented in the introduction section, this type of PTFs is beyond the scope of this work and was not tested in this study.

It is worthy to notice the important sign provided by the scatter plots of both PTFs types (point *vs.* PC) derived by various regression methods in the training phase. The fact that the tip of the data plumes (at high matric potentials) is mostly below the 1:1 reference line (left panel of Figs. 4 and 5), might indicate that in this study significant predictors of SWRC might not be optimally identified at high matric potentials. It has been widely shown that soil water retention at high matric potential is primarily determined by the soil-pore system, or in other terms 'soil structure'. Moreover, in the study of Pulido Moncada (2014), soil structure has been reported as the temporal indicator of the change of soil quality. As it is manifested with the mean test error in the present study, soil temporal variability is probably one of the hindrances of PTF's transferability. Utilising soil structure information as one of PTFs predictors is expectedly to improve PTFs' performance (Vereecken et al., 2010). Nguyen et al. (2014) recently reported that incorporating categorical soil structure information in point PTFs developed by the MLR technique improved the accuracy of the SWRC estimation of tropical paddy soils. It is interesting to further investigate whether such improved effect will still be captured by different data mining techniques and for other PTF types.

5. Conclusions

This study presented the development and validation of point and PC-PTFs to estimate SWRC of tropical delta soils from basic soil properties using various regression techniques like MLR, ANN, SVR and kNN. Evaluating the accuracy and reliability of derived PTFs asserted that all four regression techniques provide comparable accuracy in estimation of soil-water content at specific matric potentials, but the reliability of point PTFs derived by pattern recognition techniques of ANN, SVR and kNN was better than that of MLR. In case of PC-PTF, ANN generates more accurate and reliable PTFs than SVR and MLR. Although the kNN approach also performed well with the topological structure of PC-PTFs (i.e., using the logarithm of matric head as extra input variable), it is a non-parametric regression technique with a limited capacity to provide a continuous SWRC prediction.

The evaluation results confirm the superiority of the ANN and kNN approaches in modelling the relationship between soil and water as a complex system even when a limited dataset is available. These findings are significantly important for tropical delta regions, where only very few limited data are

available for PTFs development, despite the growing demand to develop soil databases in such regions. Due to the black-box and user-defined natures of ANN techniques, the practical implementation of this technique has not usually been transparent to all PTF users. The usage of the kNN method, on the other hand, would have greater benefits because of its flexibility, simplicity, accuracy and capacity to append new observations in training data sets without the need to redevelop the models again.

In order to cope with the limitation of PTFs' transferability caused by temporal variation of soil hydraulic properties, incorporating soil properties which reflect the temporal change of soil quality, e.g. soil structure, might be helpful to improve PTF accuracy and reliability. Future research about the improved effects of soil structural information on SWRC estimation in combination with different regression methods is recommended.

Acknowledgements

This research has benefitted from a statistical consult with Ghent University FIRE (Fostering Innovative Research based on Evidence). The authors highly appreciate all technical supports from Departments of Soil Science, Can Tho University and Department Soil Management, Ghent University. Special thanks to Prof. Arnout Van Messem (UGent) for the practical advises about data mining techniques.

REFERENCES

Bayat, H., Neyshaburi, M. R., Mohammadi, K., Nariman-Zadeh, N., Irannejad, M., & Gregory, A. S. (2013). Combination of artificial neural networks and fractal theory to predict soil water retention curve. *Computers and Electronics in Agriculture, 92*, 92–103. http://dx.doi.org/10.1016/j.compag.2013.01.005.

Baydaroğlu, Ö., & Koçak, K. (2014). SVR-based prediction of evaporation combined with chaotic approach. *Journal of Hydrology, 508*(0), 356–363. http://dx.doi.org/10.1016/j.jhydrol.2013.11.008.

Botula, Y. D., Nemes, A., Mafuka, P., Van Ranst, E., & Cornelis, W. M. (2013). Prediction of water retention of soils from the humid tropics by the nonparametric k-nearest neighbor approach. *Vadose Zone Journal, 12*(2), 1–17. http://dx.doi.org/10.2136/vzj2012.0123.

Botula, Y.-D., Van Ranst, E., & Cornelis, W. M. (2014). Pedotransfer functions to predict water retention for soils of the humid tropics: A review. *Revista Brasileira de Ciência do Solo, 38*, 679–698.

Brooks, R. H., & Corey, A. T. (1964). *Hydraulic properties of porous media* Hydrology Paper 3. Fort Collins, Colorado, USA: Colorado State University.

Campbell, G. S. (1974). A simple method for determining unsaturated conductivity from moisture retention data. *Soil Science, 117*, 311–314.

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*(6), 1902–1911. http://dx.doi.org/10.2136/sssaj2004.0238.

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*(3), 638–648. http://dx.doi.org/10.2136/sssaj2001.653638x.

De Vos, B., Van Meirvenne, M., Quataert, P., Deckers, J., & Muys, B. (2005). Predictive quality of pedotransfer functions for estimating bulk density of forest soils. *Soil Science Society of America Journal, 69*(2), 500–510. http://dx.doi.org/10.2136/sssaj2005.0500.

Demuth, H., & Beale, M. (2000). *Neural network toolbox*. Mathworks, Inc.

Ebrahimi, E., Bayat, H., Neyshaburi, M. R., & Zare Abyaneh, H. (2013). Prediction capability of different soil water retention curve models using artificial neural networks. *Archives of Agronomy and Soil Science, 60*(6), 859–879. http://dx.doi.org/10.1080/03650340.2013.837219.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman and Hall.

Gee, G. W., & Bauder, J. W. (1986). Particle-size analysis. In A. Klute (Ed.), *Methods of soil analysis. Part 1. Physical and Mineralogical methods* (pp. 383–411). Madison, Wisconsin, USA: American Society of Agronomy.

van Genuchten, M. T. (1980). A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal, 44*(5), 892–898. http://dx.doi.org/10.2136/sssaj1980.03615995004400050002x.

Grossman, R. B., & Reinsch, T. G. (2002). Bulk density and linear extensibility. In J. H. Dane, & G. C. Topp (Eds.), *Methods of soil Analysis: Physical methods, Part 4* (pp. 201–228). Madison, USA: Soil Science Society of America.

Gupta, S. C., & Larson, W. E. (1979). Estimating soil water retention characteristics from particle size distribution, organic matter percent, and bulk density. *Water Resources Research, 15*(6), 1633–1635.

Haghverdi, A., Cornelis, W. M., & Ghahraman, B. (2012). A pseudo-continuous neural network approach for developing water retention pedotransfer functions with limited data. *Journal of Hydrology, 442–443*, 46–54. http://dx.doi.org/10.1016/j.jhydrol.2012.03.036.

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.

Haghverdi, A., Öztürk, H. S., & Cornelis, W. M. (2014). Revisiting the pseudo continuous pedotransfer function concept: Impact of data quality and data mining method. *Geoderma, 226–227*, 31–38. http://dx.doi.org/10.1016/j.geoderma.2014.02.026.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.

Hodnett, M. G., & Tomasella, J. (2002). Marked differences between van Genuchten soil water-retention parameters for temperate and tropical soils: A new water-retention pedotransfer functions developed for tropical soils. *Geoderma, 108*(3–4), 155–180. http://dx.doi.org/10.1016/S0016-7061(02)00105-2.

Hong, W.-C., & Pai, P.-F. (2007). Potential assessment of the support vector regression technique in rainfall forecasting. *Water Resources Management, 21*(2), 495–513. http://dx.doi.org/10.1007/s11269-006-9026-2.

Hu, J., Liu, J., Liu, Y., & Gao, C. (2011). EMD-KNN model for annual average rainfall forecasting. *Journal of Hydrologic Engineering, 18*(11), 1450–1457. http://dx.doi.org/10.1061/(asce)he.1943-5584.0000481.

IUSS Working Group WRB.. (2014). *World Reference Base for Soil Resources 2014. International soil classification system for naming soils and creating legends for soil maps*. World Soil Resources Reports No. 106 (2nd ed.). Rome: FAO.

Khlosi, M., Alhamdoosh, M., Douaik, A., Gabriels, D., & Cornelis, W. M. (2016). Enhanced pedotransfer functions with

support vector machines to predict water retention of calcareous soil. *European Journal of Soil Science, 67*(3), 276–284. http://dx.doi.org/10.1111/ejss.12345.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5 ed., Vol. 6). United States: McGraw-Hill New York.

Lamorski, K., Pachepsky, Y., Sławiński, 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*(5), 1243–1247. http://dx.doi.org/10.2136/sssaj2007.0280N.

Le, V. K. (2003). *Physical fertility of typical Mekong delta soils (Vietnam) and land suitability assessment for alternative crops with rice cultivation*. Ph.D Dissertation. Ghent University: Ghent University.

Linh, T. B., Sleutel, S., Elsacker, S. V., Guong, V. T., Khoa, L. V., & Cornelis, W. M. (2015). Inclusion of upland crops in rice-based rotations affects chemical properties of clay soil. *Soil Use and Management, 31*(2), 313–320. http://dx.doi.org/10.1111/sum.12174.

Manrique, L. A., Jones, C. A., & Dyke, P. T. (1991). Predicting soil water retention characteristics from soil physical and chemical properties. *Communications in Soil Science and Plant Analysis, 22*(17–18), 1847–1860. http://dx.doi.org/10.1080/00103629109368540.

Mayr, T., & Jarvis, N. J. (1999). Pedotransfer functions to estimate soil water retention parameters for a modified Brooks–Corey type model. *Geoderma, 91*(1–2), 1–9. http://dx.doi.org/10.1016/S0016-7061(98)00129-3.

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 Research, 90*(1–2), 108–116. http://dx.doi.org/10.1016/j.still.2005.08.011.

Minasny, B., & Hartemink, A. E. (2011). Predicting soil properties in the tropics. *Earth-Science Reviews, 106*(1–2), 52–62. http://dx.doi.org/10.1016/j.earscirev.2011.01.005.

Mucherino, A., Papajorgji, P. J., & Pardalos, P. M. (2009). Data mining in agriculture. In Panos M. Pardalos, & D.-Z. Du (Eds.), *Springer optimization and its applications* (Vol. 34, p. 272). New York: Springer. http://dx.doi.org/10.1007/978-0-387-88615-2.

Nemes, A., Rawls, W. J., & Pachepsky, Y. A. (2006). Use of the nonparametric nearest neighbor approach to estimate soil hydraulic properties. *Soil Science Society of America Journal, 70*(2), 327–336. http://dx.doi.org/10.2136/sssaj2005.0128.

Nemes, A., Rawls, W. J., Pachepsky, Y. A., & van Genuchten, M. T. (2006). Sensitivity analysis of the nonparametric nearest neighbor technique to estimate soil water retention. *Vadose Zone Journal, 5*(4), 1222–1235. http://dx.doi.org/10.2136/vzj2006.0017.

Nguyen, P. M., Le, K. V., & Cornelis, W. M. (2014). Using categorical soil structure information to improve soil water retention estimates of tropical delta soils. *Soil Research, 52*(5), 443–452. http://dx.doi.org/10.1071/SR13256.

Nguyen, P. M., Le, K. V., Botula, Y.-D., & Cornelis, W. M. (2015). Evaluation of soil water retention pedotransfer functions for Vietnamese Mekong Delta soils. *Agricultural Water Management, 158*(0), 126–138. http://dx.doi.org/10.1016/j.agwat.2015.04.011.

Or, D., & Wraith, J. M. (2002). Soil water content and water potential relationships. In A. W. Warrick (Ed.), *Soil physics companion* (pp. 49–82). Boca Raton, FL: CRC Press.

Pachepsky, Y., Rajkai, K., & Tóth, B. (2015). Pedotransfer in soil physics: Trends and outlook — a review —. *Agrokémia és Talajtan, 64*(2), 339–360. http://dx.doi.org/10.1556/0088.2015.64.2.3.

Pachepsky, Y., & Rawls, W. J. (2004). *Development of pedotransfer functions in soil hydrology* (Vol. 30). Elsevier.

Pachepsky, Y. A., Rawls, W. J., & Lin, H. S. (2006). Hydropedology and pedotransfer functions. *Geoderma, 131*(3–4), 308–316. http://dx.doi.org/10.1016/j.geoderma.2005.03.012.

Pachepsky, Y. A., Rawls, W. J., & Timlin, D. J. (2013). The current status of pedotransfer functions: Their accuracy, reliability, and utility in field- and regional-scale modeling. In D. L. Corwin, K. Loague, & T. R. Ellsworth (Eds.), *Assessment of non-point source pollution in the vadose zone* (pp. 223–234). American Geophysical Union.

Pachepsky, Y., & Schaap, M. G. (2004). Data mining and exploration techniques. In Y. Pachepsky, & W. J. Rawls (Eds.), *Vol. 30. Developments in soil science* (pp. 21–32). Elsevier.

Pachepsky, Y. A., Timlin, D., & Varallyay, G. (1996). Artificial neural networks to estimate soil water retention from easily measurable data. *Soil Science Society of America Journal, 60*(3), 727–733. http://dx.doi.org/10.2136/sssaj1996.03615995006000030007x.

Patil, N., Tiwary, P., Pal, D., Bhattacharyya, T., Sarkar, D., Mandal, C., … Dongre, V. (2013). Soil water retention characteristics of black soils of India and pedotransfer functions using different approaches. *Journal of Irrigation and Drainage Engineering, 139*(4), 313–324. http://dx.doi.org/10.1061/(ASCE)IR.1943-4774.0000527.

Perkins, K., & Nimmo, J. (2009). High-quality unsaturated zone hydraulic property data for hydrologic applications. *Water Resources Research, 45*(7), W07417. http://dx.doi.org/10.1029/2008wr007497.

Pulido Moncada, M. (2014). *Integrated assessment of soil structural quality*. PhD Dissertation. Gent, Belgium: Ghent Univesity.

R Core Team. (2014). *R: A language and environment for statistical computing*. Viena, Austria: R Foundation for Statistical Computing.

Saxton, K. E., & Rawls, W. J. (2006). Soil water characteristic estimates by texture and organic matter for hydrologic solutions. *Soil Science Society of America Journal, 70*(5), 1569–1578. http://dx.doi.org/10.2136/sssaj2005.0117.

Schaap, M. G., & Leij, F. J. (1998). Using neural networks to predict soil water retention and soil hydraulic conductivity. *Soil and Tillage Research, 47*(1–2), 37–42. http://dx.doi.org/10.1016/S0167-1987(98)00070-1.

Schaap, M. G., Leij, F. J., & van Genuchten, M. T. (2001). Rosetta: A computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. *Journal of Hydrology, 251*(3–4), 163–176. http://dx.doi.org/10.1016/S0022-1694(01)00466-8.

Smola, A., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing, 14*(3), 199–222. http://dx.doi.org/10.1023/b:stco.0000035301.49549.88.

Soil Survey Staff. (1975). *Soil Taxonomy* agriculture handbook. Washington, DC., USA: Soil Conservation Service, USDA.

Tomasella, J., Hodnett, M. G., & Rossato, L. (2000). Pedotransfer functions for the estimation of soil water retention in Brazilian soils. *Soil Science Society of America Journal, 64*(1), 327–338. http://dx.doi.org/10.2136/sssaj2000.641327x.

Tomasella, J., Pachepsky, Y., Crestana, S., & Rawls, W. J. (2003). Comparison of two techniques to develop pedotransfer functions for water retention. *Soil Science Society of America Journal, 67*(4), 1085–1092. http://dx.doi.org/10.2136/sssaj2003.1085.

Twarakavi, N. K. C., Šimůnek, 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*(5), 1443–1452. http://dx.doi.org/10.2136/sssaj2008.0021.

Valipour, M., Banihabib, M. E., & Behbahani, S. M. R. (2012). Monthly inflow forecasting using autoregressive artificial neural network. *Journal of Applied Sciences, 12*(20), 2139–2147.

Valipour, M., Banihabib, M. E., & Behbahani, S. M. R. (2013). Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir. *Journal of Hydrology, 476*(0), 433–441. http://dx.doi.org/10.1016/j.jhydrol.2012.11.017.

Vereecken, H., Weynants, M., Javaux, M., Pachepsky, Y., Schaap, M. G., & Genuchten, M. T. v (2010). Using pedotransfer functions to estimate the van Genuchten–Mualem soil hydraulic properties: A review. *Vadose Zone Journal, 9*(4), 795–820. http://dx.doi.org/10.2136/vzj2010.0045.

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 Science, 37*(1), 29–38.

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*(3–4), 123–150. http://dx.doi.org/10.1016/S0022-1694(01)00464-4.