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
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:

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

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:

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

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