Journal of Hydrology 525 (2015) 598–606

Contents lists available at ScienceDirect

Journal of Hydrology

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

Impact of regression methods on improved effects of soil structure on soil water retention estimates

Phuong Minh Nguyen [a,b,*], Jan De Pue [a], Khoa Van Le [b], Wim Cornelis [a]

[a] Department of Soil Management, Faculty of Bioscience Engineering, Ghent University, 653 Coupure Links, 9000 Gent, Belgium
[b] Department of Soil Science, Faculty of Agriculture and Applied Biology, Can Tho University, Campus II, 3/2 Street, Ninh Kieu District, Can Tho City, Viet Nam

ARTICLE INFO

Article history:
Received 27 January 2015
Received in revised form 8 April 2015
Accepted 9 April 2015
Available online 18 April 2015
This manuscript was handled by Geoff Syme, Editor-in-Chief

Keywords:
Pedotransfer function
Soil water retention characteristics
Support Vector Machines
k-Nearest Neighbors

SUMMARY

Increasing the accuracy of pedotransfer functions (PTFs), an indirect method for predicting non-readily available soil features such as soil water retention characteristics (SWRC), is of crucial importance for large scale agro-hydrological modeling. Adding significant predictors (i.e., soil structure), and implementing more flexible regression algorithms are among the main strategies of PTFs improvement. The aim of this study was to investigate whether the improved effect of categorical soil structure information on estimating soil-water content at various matric potentials, which has been reported in literature, could be enduringly captured by regression techniques other than the usually applied linear regression. Two data mining techniques, i.e., Support Vector Machines (SVM), and k-Nearest Neighbors (kNN), which have been recently introduced as promising tools for PTF development, were utilized to test if the incorporation of soil structure will improve PTF's accuracy under a context of rather limited training data. The results show that incorporating descriptive soil structure information, i.e., massive, structured and structureless, as grouping criterion can improve the accuracy of PTFs derived by SVM approach in the range of matric potential of −6 to −33 kPa (average RMSE decreased up to 0.005 m^3 m^{-3} after grouping, depending on matric potentials). The improvement was primarily attributed to the outperformance of SVM-PTFs calibrated on structureless soils. No improvement was obtained with kNN technique, at least not in our study in which the data set became limited in size after grouping. Since there is an impact of regression techniques on the improved effect of incorporating qualitative soil structure information, selecting a proper technique will help to maximize the combined influence of flexible regression algorithms and soil structure information on PTF accuracy.

1. Introduction

Accurate prediction of soil water retention characteristics (SWRC) is of crucial importance for solving many soil water management problems related to agricultural, hydrological and environmental issues (Botula et al., 2012). Highly spatial and temporal variability of this soil hydraulic property together with cumbersome and expensive measurement methods make it difficult to characterize the SWRC, especially for large scale modeling (Romano and Santini, 1997). In order to get the information of SWRC in a cost-effective way, indirect estimation of such property from easily measurable or readily available soil data by using pedotransfer functions (PTFs) is becoming increasingly popular. Since the indirect estimates are often tainted with considerable uncertainty, many attempts have been devoted to improve the accuracy of these predictive functions. Pachepsky et al. (2013) stated that improvements in PTFs' predictability can be achieved by using more flexible PTF algorithms, adding more significant predictors into PTFs development, and preliminary grouping of soils.

It is widely known that SWRC is a function of soil structure and soil texture (Or and Wraith, 2002). Incorporating soil structure information into texture-based PTFs, therefore, has been reported to improve the accuracy of SWRC-PTFs (Pachepsky et al., 2006). The importance of soil structure on SWRC estimation has been demonstrated in many researches (Abbaspour and Moon, 1992; Danalatos et al., 1994; Kay and Angers, 2002; McKenzie and MacLeod, 1989; Nguyen et al., 2014; Pachepsky et al., 2006; Rawls and Pachepsky, 2002; Williams et al., 1983). In a recent review on using PTFs for estimating hydraulic parameters, Vereecken et al. (2010) concluded that further improvement of PTFs is mainly limited by a lack of new information such as soil

* Corresponding author at: Department of Soil Management, Faculty of Bioscience Engineering, Ghent University, 653 Coupure Links, 9000 Gent, Belgium. Tel.: +32 9 264 60 39; fax: +32 9 264 62 47.
E-mail addresses: MinhPhuong.Nguyen@ugent.be, nmphuong@ctu.edu.vn (P.M. Nguyen).

http://dx.doi.org/10.1016/j.jhydrol.2015.04.014
0022-1694/

structure and questioned how to best include soil structural information into PTFs. Depending on how soil structure characteristics are assessed (e.g., visual morphology description or quantitative soil structural indices), soil structure information can be incorporated directly as PTF's input predictor (Giménez et al., 2001; Lin et al., 1999; Pachepsky et al., 1998) or as grouping criterion to partition soils into homogeneous subgroups with similar SWRC for specific PTF development (Danalatos et al., 1994; Nguyen et al., 2014; Pachepsky and Rawls, 2003; Pachepsky et al., 2006; Rawls and Pachepsky, 2002; Williams et al., 1992).

In the literature, most of the studies about the effect of soil structure on the accuracy of SWRC-PTFs were carried out via statistical regression techniques – the main stay of statistics and still remaining the most important tool for prediction problems (Hastie et al., 2009). With the increasing popularity of data-mining techniques nowadays as a powerful tool for PTF development, it is very useful to examine the impact of various regression methods on the PTF accuracy when categorical soil structure information that is routinely available from soil survey databases, is incorporated into PTFs development. Although other indirect properties to express soil structure such as penetration resistance (Pachepsky et al., 1998), soil aggregation or those extracted from recent developments, like x-ray tomography, spectral induced polarization, and nuclear magnetic resonance (Vereecken et al., 2010), might potentially improve PTF predictions, they are simply not available in most databases.

Support Vector Machines (SVM) and k-Nearest Neighbors (kNN) are two promising approaches employed in PTF research nowadays, due to their flexibility and accurate performance (Botula et al., 2013). The kNN approach was applied successfully to develop SWRC-PTFs for soils in both temperate (Nemes et al., 2006a) and tropical regions (Botula et al., 2013). Based upon the good results of such PTFs, the user-friendly 'k-Nearest' software was introduced by Nemes et al. (2008) to estimate soil water content at field capacity (FC) and permanent wilting point (PWP) based on kNN algorithm. SVM is now considered as a promising alternative to artificial neural networks (ANN) as it can eliminate the local minimum issue, which is the main weakness of the ANN approach. Lamorski et al. (2008) applied SVM to predict soil water content at various pressure heads using simple basic properties of Polish soils. A similar SVM approach was employed by Twarakavi et al. (2009) to develop parameter-based PTFs for van Genuchten–Mualem models. The SVM approach has not yet been applied for soils in the humid tropics (Botula et al., 2014).

It is widely known that the lack of well-defined and extensive data of soil hydraulic properties in the tropics is one of the major limitations dragging the development of SWRC-PTFs behind (Hodnett and Tomasella, 2002), although the need of accurate and up-to-date information of soil hydraulic properties in such regions is even more urgent than elsewhere as they are sparse and outdated (Minasny and Hartemink, 2011). Nowadays, many attempts have been devoted to study soil-water relationships of tropical soils through developing SWRC-PTFs. These PTFs have been derived using rather limited data which represented specific soils in tropical regions; for instance, highly weathered soils on stable landforms (Botula, 2013), recently developed alluvial soils in a dynamic river basin (Nguyen et al., 2014), and black clayey soils with strong shrinking and swelling characteristics (Patil et al., 2013).

Based upon the above, it is clear that there is a need to explore the interactions of different strategies on the improvement of PTF's accuracy when only limited data sets are available for PTF development. Since each regression approach has its own situation to work best (Hastie et al., 2009), the objective of this study was to investigate whether incorporating categorical soil structure information will improve the accuracy of PTFs developed based on SVM and kNN approaches. This study contributes in addressing the question on how to best include soil structural information into PTFs by testing different combinations of flexible regression approaches with considering categorical soil structural information. It would in turn be valuable to those interested in developing new PTFs with rather limited data sets at hand.

2. Materials and methods

2.1. Soil data set

A data set of 160 samples from tropical Vietnamese Mekong Delta (VMD) which was used to derive the Multiple Linear Regression (MLR) PTFs by Nguyen et al. (2014) was utilized in this study. The data includes the records of morphological and physico-chemical soil properties of two upper diagnostic horizons taken from agricultural fields (mainly paddy rice, but also upland crops such as sugarcane, maize and watermelon). The thickness of the two upper horizons varies from site to site with maximum lower boundary of 25 cm for surface horizons and 70 cm for subsurface horizons. The soils in the study area are classified as Fluvisols, Gleysols, Luvisols, Acrisols, Arenosols and Plinthosols (IUSS Working Group WRB, 2014).

Undisturbed soil samples, taken by standard sharpened steel cylinders of 100 cm^3, were used to determine soil bulk density (BD), by core method (Grossman and Reinsch, 2002) and SWRC at eight matric potentials (e.g., −1, −3, −6, −10, 20, −33, −100, −1500 kPa) using sandbox apparatus and pressure chambers, according to the procedures outlined in Cornelis et al. (2005). Disturbed soil samples taken nearby the place of undisturbed sampling were used to determine organic carbon content by wet oxidation method (Walkley and Black, 1934) and particle size distribution by sieve-pipette method (Gee and Bauder, 1986).

The morphological soil structure description in terms of types of soil structure according to the FAO Guidelines for Soil Description (FAO, 2006) was used as grouping criterion to partition data into three uniform subsets of massive, structured and structureless soils with sample size of 91, 46 and 23, respectively. The main reason of using only aspect of soil structure type is the size of the available training data set, which is rather small. Other information of soil structure description such as grade of soil structure development, size and shape of structural units are not considered in the present study since further grouping soils by detailed structural information can lead to more homogeneous, yet small subsets with only few observations which in turn are not adequate for PTF development. The potential power of these soil structural aspects on SWRC estimates should be given due consideration when a large database of tropical delta soils becomes available in the future.

In the present study, the basic soil properties (i.e., sand, silt, clay content, organic carbon content (OC), and bulk density (BD) were utilized as predictors for SWRC estimation with both SVM and kNN techniques. Descriptive statistics of predictor variables and SWRC of the whole data set and the three structural subsets are presented in Table 1. The variation of soil texture in the data set is graphically displayed in Fig. 1, where different markers represent the samples of different structural subgroups.

2.2. Methods to build PTFs

2.2.1. Support Vector Machines

Support Vector Machines (SVM) proposed by Vapnik (1995) is statistical learning machines with a promising ability to generalize the prediction. Instead of minimizing the observed training error as in statistical regression techniques, SVM attempts to minimize the