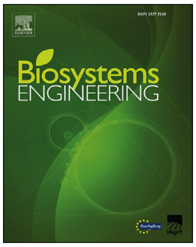
Available online at www.sciencedirect.com

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

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