Journal of Hydrology 549 (2017) 592–602

Contents lists available at ScienceDirect

Journal of Hydrology

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

Research papers

Modeling soil bulk density through a complete data scanning procedure: Heuristic alternatives

Jalal Shiri [a], Ali Keshavarzi [b,*], Ozgur Kisi [c], Sepideh Karimi [a], Ursula Iturraran-Viveros [d]

[a] Water Engineering Department, Faculty of Agriculture, University of Tabriz, Tabriz, Iran
[b] Laboratory of Remote Sensing and GIS, Department of Soil Science, University of Tehran, P.O. Box: 4111, Karaj 31587-77871, Iran
[c] Center for Interdisciplinary Research, International Black Sea University, Tbilisi, Georgia
[d] Departamento de Matematicas, Facultad de Ciencias, Universidad Nacional Autonoma de Mexico, Circuito Escolar, Cd. Universitaria, Coyoacán 04510, Ciudad de México, Mexico

ARTICLE INFO

Article history:
Received 28 February 2017
Received in revised form 13 April 2017
Accepted 15 April 2017
Available online 19 April 2017

This manuscript was handled by G. Syme, Editor-in-Chief

Keywords:
Heuristic models
K-fold testing
Pedotransfer functions
Soil bulk density

ABSTRACT

Soil bulk density (BD) is very important factor in land drainage and reclamation, irrigation scheduling (for estimating the soil volumetric water content), and assessing soil carbon and nutrient stock as well as determining the pollutant mass balance in soils. Numerous pedotransfer functions have been suggested so far to relate the soil BD values to soil parameters (e.g. soil separates, carbon content, etc). The present paper aims at simulating soil BD using easily measured soil variables through heuristic gene expression programming (GEP), neural networks (NN), random forest (RF), support vector machine (SVM), and boosted regression trees (BT) techniques. The statistical Gamma test was utilized to identify the most influential soil parameters on BD. The applied models were assessed through k-fold testing where all the available data patterns were involved in the both training and testing stages, which provide an accurate assessment of the models accuracy. Some existing pedotransfer functions were also applied and compared with the heuristic models. The obtained results revealed that the heuristic GEP model outperformed the other applied models globally and per test stage. Nevertheless, the performance accuracy of the applied heuristic models was much better than those of the applied pedotransfer functions. Using k-fold testing provides a more-in-detail judgment of the models.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

Soil bulk density (BD) is defined as the dry weight of soil per soil volume (which includes the soil particles as well as the soil pores). It is an important factor in land drainage and reclamation because it is an indicator of drainage characteristics (Arya and Paris, 1981; Braun and Kruijne, 1994), and determines whether there are impermeable barriers in the soil, which can deteriorate the drainage and root penetration conditions (Lampurlanes and Cantero-Martinez, 2003). In irrigation scheduling, BD can be utilized to estimate the soil volumetric water content which is an important parameter for controlling optimum irrigation (Howell and Meron, 2007). BD is also an essential factor for assessing soil carbon and nutrient stock (Ellert and Bettany, 1995), determining pollutant mass balance in soil, and determining the soils' packing structure in soil classification issues (Dexter, 1988). It also affects the soil biomass productivity and environment quality (Lal and Kimble, 2001).

Recently, the soil processes simulating models have been developed for improving the existing knowledge on important soil processes as well as evaluating the agricultural and environmental problems (Minasny and McBratney, 2002). On the other hand, soil properties continuously vary across the landscape. Nevertheless, there are usually limited numbers of direct observations (visual inspection, sampling and observation) in the field, which make it difficult to determine some soil properties directly (Heuvelink and Webster, 2001). So, numerous investigations have been conducted to relate some soil properties to easily measured available soil characteristics (e.g. particle-size distribution, organic matter or organic C content, BD, porosity, etc). Such relationships are called as pedotransfer functions (Mermoud and Xu, 2006).

The common methods of developing pedotransfer functions are multi variate-linear regression and artificial intelligence-based models (Schaap and Leij, 1998). Among others, Jalabert et al. (2010) applied boosted regression trees (BT) for estimating forest soil BD and found that the variations in forest soil BD magnitudes are largely affected by organic carbon (OC) content, followed by tree species, the coarse fragment content, parent material and depth of sampling. Ghehi et al. (2012) applied k-nearest neighbor

* Corresponding author.
E-mail address: alikeshavarzi@ut.ac.ir (A. Keshavarzi).

http://dx.doi.org/10.1016/j.jhydrol.2017.04.035
0022-1694/© 2017 Elsevier B.V. All rights reserved.

and boosted trees methods for predicting top soil BD of a tropical mountain forest soils and found the OC as the most influential parameter on BD. Al-Qinna and Jaber (2013) used different techniques inducing linear/nonlinear regression and neural networks (NN) for estimating BD using data from an arid environment in Jordan and found NN as the best model among other applied models. Botula et al. (2015) compared multivariate linear regression and k-nearest neighbor methods for predicting BD for the soils of Central Africa. They used independent soil samples for further testing of the developed models and observed substantial differences between the observed and predicted BD magnitudes, which imply the difficulties in generalizing the pedotransfer functions for its estimation. Rodríguez-Lado et al. (2015) compared random forest (RF), linear regression and NN in estimating BD and found the RF as the most accurate model among the applied models. They also confirmed that the soil BD in the studied area was mainly influenced by the soil OC. Xiangsheng et al. (2016) compared linear regression and NN models to develop pedotransfer functions for BD estimation and confirmed the superiority of NN, which confirms the conclusions obtained by Patil and Chaturvedi (2012) regarding the NN superiority.

The presented studies have used different methods for determining the input variables of the applied models. In heuristic models, feature selection (or variable choice) is the way toward choosing a subset of important elements for utilizing in model building. The focal preface when utilizing a feature selection strategy is that the data involves various features that are unessential, and can in this manner be expelled without bringing about much loss of information. Unessential features are two particular concepts, since one pertinent feature might be repetitive in the asset of another appropriate feature with which it is highly correlated. There are two types of variable selection methods: wrapper and filter types (Pfleger et al., 1994). Wrapper techniques assess variable subsets which permit to identify the probable interactions between variables (see Phuong et al., 2005). Filter methods (e.g. Gamma test, which was used in the current study) evaluate the set of variables directly from the data itself. Model construction based on wrapper methods is rather inflexible though it may be advantageous in some situations where model selection is integrated together with the variable selection method.

Nonetheless, most of the existing literatures have used a single data set assignment for developing and testing the applied regression and heuristic models, where a part of available patterns are used for training the models, then the obtained models are tested using the rest of available patterns. Meanwhile, some studies have tried to test the obtained models using data outside the studied region as discussed by Botula et al. (2015). However, the present paper focused on developing pedotransfer functions of soil BD estimation by using gene expression programming (GEP), neural networks (NN), random forest (RF), support vector machine (SVM) and boosted regression trees (BT) techniques assessed through a k-fold testing. So, a complete data set scanning was conducted so that all available patterns can participate in train-test phases, which will avoid obtaining partially valid conclusions (that might be achieved using single data set assignment). A comparison was also made between the results of these models and those obtained through using the previously suggested regression-based pedotransfer functions.

2. Study area and used data

Data used in the present study were gathered in Mohr plain, Fars province, located in Southwestern Iran (Fig. 1), between the latitudes of 27° 25′ N to 27° 59′ N and longitudes of 52° 21′ E to 53° 05′ E, with an area about 1900 km^2. The area altitude varies from 282 m to 1780 m, while the main topographic elevation ranges over 1031 m (above sea level). The dominant soil types include Lithic Leptosols, Gypsic Regosols, Haplic Calsisols, Calcaric Cambisols and Calcaric Solonchaks, which cover mountains, hilly land, plains and colluvial fans.

The main land uses practiced in the study area are pastures and irrigated farming across the Mehran River. A simple random sampling scheme was designed using ArcGIS 10.2.2 software for an appropriate determination of soil sampling areas to consider spatial varieties of the parameters affecting the BD in the study region. A total of 250 soil samples were obtained from two-first vertical depths (0–30 and 30–60 cm depth) of 125 representative soil profiles. Depths were assigned to a soil textural class determined by the substances of clay, silt, and sand, as indicated by the USDA textural triangle (Schoeneberger et al., 2002). The soil texture classes are illustrated in Fig. 2. Table 1 summarizes the statistical characteristics of the applied data. From Table 1 it is seen that soil BD has a wide variability range with the minimum and maximum values of 1.134 and 1.964 g/cm^3, respectively. The variations of the other applied parameters (except pH) are also strongly high representing high variability class according to the coefficient of variation values (Adrover et al., 2012).

The sampling sites were designed to cover equally the entire area and to incorporate different soil and land use types. The collected disturbed soil samples were air dried, crushed and sieved using 2 mm sieve size. Large plant material and pebbles were separated and discarded. Soil organic carbon (OC) content was obtained by the Walkley–Black technique with dichromate extraction and titrimetric quantization (Nelson and Sommers, 1986). Rates of clay (<0.002 mm), silt (0.002–0.05 mm), and sand (0.05–2 mm) particles were measured via hydrometer method (Gee and Bauder, 1986). Soil pH was measured in saturated paste extract utilizing a digital pH-meter (Thomas, 1996) and calcium carbonate equivalent (CCE) was obtained by the back-titration technique (Nelson, 1982). Finally, the clod method (Blake and Hartge, 1986) was utilized for determining BD with triple replications.

2.1. Input selection

The Gamma test is a filter method that assesses the significance of the variables directly from the data. A rigorous mathematical proof of the method is presented in the work by Evans (2002). The Gamma test was originally conceived as a method of estimating the variance of the residual of a model. It has also been used successfully in the estimation of optimum embedding dimension and lag of chaotic systems (Tsui et al., 2002) and the assessment of the quality of data (Jones et al., 2002). The Gamma test can be used to predict the reliability of models before the heuristic models training phase begins saving lots of time. In this paper, we seek to estimate the BD from 6 input variables: Clay, Silt, Sand, CCE, OC and pH. Therefore, the basic relationship of the system under investigation is of the accompanying form:

$$Bulk\quad density = f(\mathrm{Clay},\ \mathrm{Silt},\ \mathrm{Sand},\ \mathrm{CCE},\ \mathrm{OC},\ \mathrm{pH}) + r, \tag{1}$$

where f is a smooth function and r is a random variable that indicates noise. The domain of conceivable models is limited to the class of smooth capacities which have limited first partial derivatives. The Gamma statistic Γ is the estimate of that part of the output variance that can't be represented by a smooth data model.

Taking into account that we consider 6 possible input variables for the model building then the number of possible combinations is $2^6 - 1 = 63$. Among these possible input combinations, we might want to choose those which have the least Γ statistic. The Gamma statistic Γ also gives a lower bound for the mean square error (MSE). When training heuristic models, this implies that we should stop the training process when the MSE reaches the Gamma

Download English Version:

https://daneshyari.com/en/article/5771008

Download Persian Version:

https://daneshyari.com/article/5771008

Daneshyari.com