Geoderma 213 (2014) 64–73

Contents lists available at ScienceDirect

Geoderma

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

Predicting soil bulk density for incomplete databases

Cleiton H. Sequeira [a,b,*], Skye A. Wills [b], Cathy A. Seybold [b], Larry T. West [b]

[a] School of Natural Resources, University of Nebraska-Lincoln, Hardin Hall 3310 Holdrege Street, Lincoln, NE 68583-0961, USA
[b] USDA-NRCS National Soil Survey Center, 100 Centennial Mall North, Room 152, Lincoln, NE 68508-3866, USA

ARTICLE INFO

Article history:
Received 8 March 2012
Received in revised form 9 June 2013
Accepted 21 July 2013
Available online 7 September 2013

Keywords:
Bulk density
Pedotranfer functions
Random forest
Pedon description

ABSTRACT

Soil bulk density (ρ_b) is important because of its direct effect on soil properties (e.g., porosity, soil moisture availability) and crop yield. Additionally, ρ_b measurements are needed to express soil organic carbon (SOC) and other nutrient stocks on an area basis (kg ha^{-1}). However, ρ_b measurements are commonly missing from databases for reasons that include omission due to sampling constraints and laboratory mishandling. The objective of this study was to investigate the performance of novel pedotransfer functions (PTFs) in predicting ρ_b as a function of textural class and basic pedon description information extracted from the horizon of interest (the horizon for which ρ_b is being predicted), and ρ_b, textural class, and basic pedon description information extracted from horizons above or below and directly adjacent or not adjacent to the horizon of interest. A total of 2,680 pedons (20,045 horizons) were gathered from the USDA-NRCS National Soil Survey Center characterization database. Twelve ρ_b PTFs were developed by combining PTF types, database configurations, and horizon limiting depths. Different PTF types were created considering the direction of prediction in the soil profile: upward and downward prediction models. Multiple database configurations were used to mimic different scenarios of horizons missing ρ_b values: random missing (e.g., ρ_b sample lost in transit) and patterned or systematic missing (e.g., no ρ_b samples collected for horizons > 30 cm depth). For each database configuration scenario, upward and downward models were developed separately. Three limiting depths (20, 30, and 50 cm) were tested to identify any threshold depth between upward and downward models. For both PTF types, validation results indicated that models derived from the database configuration mimicking random horizons missing ρ_b performed better than those derived from the configuration mimicking clear patterns of missing ρ_b measurements. All 12 PTFs performed well (RMSPE: 0.10–0.15 g cm^{-3}). The threshold depth of 50 cm most successfully split the database between upward and downward models. For all PTFs, the ρ_b of other horizons in the soil profile was the most important variable in predicting ρ_b. The proposed PTFs provide reasonably accurate ρ_b predictions, and have the potential to help researchers and other users to fill gaps in their database without complicated data acquisition.

1. Introduction

Soil bulk density (ρ_b) is important because of its direct effect on soil properties such as porosity, soil moisture availability, and hydraulic conductivity (Dam et al., 2005), and its indirect effects on root growth and crop yield (Reichert et al., 2009). In addition to these well defined physical and biological roles, ρ_b measurements are needed to convert soil organic carbon (SOC) and other nutrient stocks, at any specified depth, from a mass basis (g kg^{-1}) to an area basis (kg ha^{-1}). In SOC stock studies, calculations on an area basis are preferred in order to account for differences in ρ_b with land use or management change (Ellert and Bettany, 1995; Gál et al., 2007; Verhulst et al., 2010).

Despite the importance of ρ_b, it is relatively common to find databases or datasets worldwide that are lacking ρ_b measurements for all or some records. One common reason for this is that ρ_b measurements are labor intensive, time-consuming, and expensive. A need for ρ_b data has led to the development of a variety of pedotransfer functions (PTFs) that predict ρ_b using information from more easily obtainable and available data, including physical and chemical soil properties such as soil texture, SOC, pH, and exchangeable cations (Adams, 1973; De Vos et al., 2005; Heuscher et al., 2005; Rawls, 1983), as well as morphology and landscape information such as parental material, horizon designation, physiography, and vegetation (Calhoun et al., 2001; Jalabert et al., 2010).

Previous studies have suggested stratifying databases into surface/near-surface and sub-surface samples and then developing ρ_b PTFs for

Abbreviations: CD, continuous database; DD, discontinuous database; DWM, downward model; HOI, horizon of interest; HAD, horizon with available bulk density; PTF, pedotransfer function; RF, random forest; SOC, soil organic carbon; UWM, upward model.

* Corresponding author at: USDA-NRCS National Soil Survey Center, 100 Centennial Mall North, Room 152, Lincoln, NE 68508-3866, USA. Tel.: +1 402 437 4135; fax: +1 402 437 5760.
E-mail addresses: csequeira2@unl.edu (C.H. Sequeira), skye.wills@lin.usda.gov (S.A. Wills), cathy.seybold@lin.usda.gov (C.A. Seybold), larry.west@lin.usda.gov (L.T. West).

0016-7061/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.geoderma.2013.07.013

each depth-strata. Stratification by depth accounts for the greater effect of soil management practices and plant roots on ρ_b of surface/near-surface samples than on ρ_b of sub-surface samples (Benites et al., 2007; Watson et al., 2000) and overburden pressures on sub-surface samples. Predictive accuracy also depends on the variables (e.g., soil properties, land cover) selected to predict ρ_b. It is important to select only variables that significantly affect ρ_b so that effort and resources can be optimized. The use of only relevant variables should lead to the development of simplified models that not only reduce the cost of collecting irrelevant data, but also to reduce the risk of overfitting the model, which reduces the prediction accuracy for unseen (new) data (Aertsen et al., 2010; Chan et al., 2011). In an evaluation of variable importance, Jalabert et al. (2010) found that SOC was the most important variable in the prediction of ρ_b followed by the dominant forest tree species, gravel content, parent material, depth, silt content, pH, clay content, and sum of exchangeable cations (Ca^{2+}, Mg^{2+}, and K^{+}). The importance of selected variables, however, depends on the prediction situation and modeling criteria. For instance, Benites et al. (2007) reported that SOC was more important than clay content in predicting ρ_b for the top 30 cm of soil but that the inverse was true for PTFs developed for the 30–100 cm depth due to the inverse relationship between SOC content and depth for most soils. Additionally, the algorithm chosen for fitting the PTFs also plays an important role in prediction performance. Multiple linear regression (MLR) has been the method most used for developing ρ_b PTFs; however, several non-parametric approaches such as artificial neural networks (ANN), k-nearest neighbor (k-NN), random forest (RF), and boosted regression are also viable techniques that have not been extensively used for ρ_b prediction (Jalabert et al., 2010; Tranter et al., 2007). Some of the main advantages of these non-parametric methods are being distribution-free and flexible to work with categorical variables without the need to create numerical dummy variables. One important disadvantage of these methods, however, is that they do not deliver a final equation at the end of the modeling process, making it necessary for interested users to re-fit the model for subsequent predictions. This limitation can be minimized by well documented instructions of how to recreate the model.

Most ρ_b PTFs assume that ρ_b measurements are missing from all samples in the database and that other more easily obtainable data (in the database) are available for predicting ρ_b. However, this is not always the case. There are cases in which soil samples are collected from the entire soil profile for visual, textural, and chemical characterization, but ρ_b samples are collected just from the upper soil horizons/layers due to time, budget or other constraints. In these situations, the database may consistently lack ρ_b measurements for subsurface layers or horizons. Another situation would be the loss of random ρ_b samples in the database due to data entry errors, presence of fragments, laboratory mishandling, and/or transport mishaps. In this case, ρ_b measurements would be randomly missing in the database without any specific pattern of missing ρ_b data. In both cases, however, not all ρ_b measurements are missing from the database, creating the opportunity to use existing ρ_b measurements in the prediction of missing values.

The USDA-NRCS National Soil Survey Center has, over the past several decades, collected and stored laboratory and descriptive data (pedon morphology descriptions) for the contiguous U.S. with the mission to cooperatively investigate, inventory, document, classify, interpret, disseminate, and publish information about soils of the U.S. (http://nrcs.usda.gov/). As with many other databases, however, the USDA-NRCS National Soil Survey Center's database also presents missing ρ_b measurements that can be predicted with existing ρ_b measurements and other basic information. Thus, the objective of this study was to investigate the performance of novel PTFs in predicting ρ_b as a function of textural class and basic pedon description information extracted from the horizon of interest (the horizon for which ρ_b is being predicted), and ρ_b, textural class, and basic pedon description information extracted from horizons above or below and adjacent or not adjacent to the horizon of interest.

2. Material and methods

2.1. The database

Table 1 lists some properties of the 2,680 pedons (20,045 horizons) from the contiguous U.S. that were gathered from the USDA-NRCS National Soil Survey Center characterization database (http://soils.usda.gov/survey/nscd/, accessed on 09/27/2011) and utilized in the present study for developing ρ_b PTFs. All ρ_b measurements were determined at −33 kPa matric potential using the clod method and particle-size distribution (clay, silt, and sand contents) analysis was performed by the pipet method according to Soil Survey Lab protocols (Soil Survey Staff, 2004). Horizons were assigned to a soil textural class determined by the contents of clay, silt, and sand, according to the USDA textural triangle (Schoeneberger et al., 2002). Soil OC was determined by dry combustion (total C) for samples without carbonates present and by the difference between total C and inorganic C (pressure calcimeter method) for samples with carbonates present (Soil Survey Staff, 2004). In addition to the properties presented in Table 1, the database was queried for selected pedon description information (described in the next section). The database includes horizon designations (e.g., Ap, Bt) and top and bottom depths described and recorded according to Soil Survey Staff (2004).

2.2. The random forest algorithm

Random forest, a tree-based algorithm developed by Breiman (2001), was chosen for developing the PTFs due to several characteristics. This algorithm can handle a mixture of categorical and continuous variable; can handle unbalanced classes; incorporates interactions among predictor variables; returns variable importance; requires little need to fine-tune parameters (Breiman, 2001; Izenman, 2008). It has been claimed that random forest does not overfit (Breiman, 2001) but other studies have indicated that it is not always the case (Luellen et al., 2005). The goal of RF is to obtain stable predictors (regressors) and, hence, robust models by applying two randomization procedures: bagging (bootstrap aggregating) and random input selection. The bagging procedure draws random and independent B bootstrap samples from the learning (calibration) set to grow B regression trees (Breiman, 1996). Each bootstrap sample is obtained by repeated sampling with replacement from the calibration set. In other words, there are equal probabilities on the sample points to be selected on each bootstrap sample. Random input selection randomly selects a subset of variables to determine the best split at each node in the tree (Ho, 1998). These two randomizations are crucial for obtaining stable predictors. In the present study, each random forest was grown with 1,500 trees to guarantee an accurate error rate. To reduce bias, trees were grown to maximum depths with no pruning. Following the default value, one third of the input variables were randomly sampled to determine the best split at each node. This default configuration has been reported to be a good choice (Liaw and Wiener, 2002). Random forests were developed using the randomForest package of R (R Development Core Team, 2011).

Table 1
Descriptive statistics of data used for developing bulk density models.

	N[a]	Mean	SD[b]	Minimum	Maximum
Bulk density, g cm^{-3}	20,045	1.41	0.21	0.23	2.41
Organic Carbon, g kg^{-1}	16,881	7.90	12.5	0.00	467
Clay, g kg^{-1}	20,045	265	160	0.00	931
Silt, g kg^{-1}	20,045	403	198	0.00	955
Sand, g kg^{-1}	20,045	301	245	0.00	978

[a] Sample size.
[b] Standard deviation.