2014

Predicting soil bulk density for incomplete databases

Cleiton H. Sequeira

University of Nebraska-Lincoln, csequeira2@unl.edu

Skye A. Wills

USDA-NRCS, skye.wills@lin.usda.gov

Cathy A. Seybold

USDA-NRCS, cathy.seybold@lin.usda.gov

Larry T. West

USDA-NRCS, larry.west@lin.usda.gov

Follow this and additional works at: <http://digitalcommons.unl.edu/natrespapers>

Sequeira, Cleiton H.; Wills, Skye A.; Seybold, Cathy A.; and West, Larry T., "Predicting soil bulk density for incomplete databases" (2014). *Papers in Natural Resources*. 397.

<http://digitalcommons.unl.edu/natrespapers/397>

This Article is brought to you for free and open access by the Natural Resources, School of at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Papers in Natural Resources by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

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

Pedotransfer 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.

© 2013 Elsevier B.V. All rights reserved.

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).

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.

2.3. Input variables

In general, the proposed PTFs had 9 input variables (regressors): 4 from the horizon of interest [designation, textural class, depth (at the middle of the horizon), and thickness] and 5 from a horizon that could be above or below and adjacent or not adjacent to the horizon of interest (ρ_b , horizon designation, textural class, depth, and thickness). Horizon depth was calculated as the middle of the horizon range and horizon thickness as the difference between bottom and top depths. The use of soil textural class (categorical variable) instead of the contents of clay, silt, and/or sand (continuous variable) gives the PTFs the flexibility to deal with databases where only field determined soil textural class is available (tactile determinations). It has been shown that there is a close relationship between field and lab soil textural determinations (Foss et al., 1975; Post et al., 1986). The majority of published PTFs use SOC as a predictor variable. However, SOC was not used in this study because it is not always evaluated for all horizons in the database. A similar approach was employed by Zacharias and Wessolek (2007) who excluded organic matter, and used soil texture and bulk density to predict soil water retention.

The RF algorithm can handle discrete variables with a maximum of 32 categories. Therefore, it was necessary to combine the original 511 different horizon designations present in the database into 32 or less categories. Lithologic discontinuity designations were ignored due to their unpredictable effect on ρ_b (e.g., Bt2, 2Bt, and 3Bt1 were grouped in the general category Bt). Transition horizons were generally pooled with the first described master horizon due to the usual predominance of characteristics of the first described master horizon (e.g., A/E, A/O, and AE horizon were considered A). Other horizon designations that occurred infrequently or did not connote any expected significant effect on ρ_b values were grouped by approximations (e.g., Bg and Bw horizons were included in the general category B). This approach led to the formation of the 20 horizon designation categories shown in Table 2. Box-and-whisker and scatter plots were created to visualize the relationship between observed ρ_b and input variables using the Graphics package of R (R Development Core Team, 2011).

2.4. Pedotransfer function development

Separate PTFs were developed depending on where available ρ_b measurement was located in the soil profile (below or above the

horizon of interest). In this study, PTFs were referred to as upward models (UWMs) when the horizon with available ρ_b was below or downward models (DWMs) when the horizon with available ρ_b was above the horizon of interest. The PTF names refer to the direction of prediction.

In the next step of model development, for both PTF types (UWM and DWM), the database was arranged in two different configurations to simulate two situation with different ρ_b availability. In the first configuration, which we will refer to as the continuous database (CD), the data were reconfigured to mimic a situation where random horizons are missing ρ_b measurement. In this case, the horizon with available ρ_b is the horizon directly adjacent to the horizon of interest (below it if developing UWMs and above it if developing DWMs). For a hypothetical pedon with horizons at 0–5, 5–20, 20–30, 30–50, 50–100, and 100–150 cm depth, the horizon of interest at 20–30 cm depth would use information from the adjacent horizon at 30–50 cm depth for developing a UWM, and from the adjacent horizon at 5–20 cm depth for developing a DWM. Following this arrangement, the horizon of interest at 30–50 cm depth would use information from the adjacent horizon at 50–100 cm depth for developing a UWM, and from the adjacent horizon at 20–30 cm depth for developing a DWM. Columns 3–6 of Table 3 illustrate how the CD configuration of the hypothetical pedon described above for UWMs with a maximum depth of 30 cm and for DWMs with a depth greater than 30 cm.

In the second database configuration, which we refer to as the discontinuous database (DD), the data were reconfigured to mimic a situation where there is a portion in all soil profiles where there are missing ρ_b measurements (e.g., no ρ_b measurements at depths above or below 30 cm). In this scenario, adjacent horizons are not available for every horizon of interest. For the same hypothetical pedon described above, mimicking a database that is missing ρ_b measurements for horizons above 30 cm depth, the horizon of interest at 20–30 cm depth would use information from the adjacent horizon at 30–50 cm depth for developing a UWM. However, when the horizon of interest is the one at 5–20 cm depth, it would not have ρ_b available from an adjacent horizon. In this case, the horizon at 30–50 cm depth would be used in the UWM as the closest horizon with an available ρ_b . The same horizon at 30–50 cm depth would be used when the horizon of interest is at 0–5 cm depth. For an alternate case, mimicking a situation where no ρ_b measurements were collected for horizons deeper than 30 cm, the horizon of interest at 30–50 cm depth would use information from the adjacent horizon at 20–30 cm depth for developing a DWM. However, the horizons of interest at 50–100 and 100–150 cm depth would not have ρ_b available from adjacent horizons (30–50 and 50–100 cm depth, respectively), making the horizon at 20–30 cm depth as the closest with available ρ_b . Columns 7–10 of Table 3 illustrate how the DD

Table 2

Strategy used for grouping individual horizon designations in the dataset into final 20 categories of horizon designation. Individual horizons were described according to the Field Book for Describing and Sampling Soils (Schoeneberger et al., 2002).

Final horizon designation	Individual horizon designation	N ^a
A	A, A/E, A/O, AE, Ad Ay	2420
AB	A/B, Ab, ABd, ABgb, ABt, Agb	437
Ap	Ap, Abp, Agp, Ap/C	2151
B	B, B', BE, B/E, B/C, B and E, Bg, Bj, Bn, Bw, Bwg	2896
BA	BA, B/A	4
Bd	Bd, Bad, BCd,	24
Bhs	Bhs, Bh, Bs	74
Bk	Bk, Bkb, Bkg, Bkk, Bkg, Bkny, Bkqy	1576
Bm	Bm, Bkm, Bkqm, Bsm,	24
Bq	Bq	11
Bt	Bt, Btg, Btk, Btn, BCT, Bss, Btss, Btb, Btz	6785
Bv	Btv, Bv, BCv	28
Bx	Bx, Btx, Btgx, B/Ex, BCx,	255
By	By, Byy, Byz	118
C	C, CB, C/B, Cg, Ck Ct, Cy,	2421
Cd	Cd	160
Cr	Cr	114
E	E, EB, E/B, E and B, E/Bt, E and Bt, E/A, EA, EC	535
O	Oi, Oa, Oe, O/A, L, Lco, Ldi, Lk, Lma	5
R	R	7

^a Sample size.

Table 3

Two different ways of relating information from horizon of interest (HOI) and horizon with available bulk density (HAD) in a soil profile for developing both upward models (UWMs) and downward models (DWMs) with threshold depth of 30 cm. The difference relies on data being arranged as continuous dataset (mimicking randomly missing bulk density values) or discontinuous dataset (mimicking pattern-consistent missing bulk density values). In both dataset configurations, the HOI is kept in its original position while HAD is moved.

Horizon depth	Horizon ID ^a	Continuous database				Discontinuous database			
		UWM		DWM		UWM		DWM	
		HOI	HAD	HOI	HAD	HOI	HAD	HOI	HAD
0–5	1	1	2	–	–	1	4	–	–
5–20	2	2	3	–	–	2	4	–	–
20–30	3	3	4	–	–	3	4	–	–
30–50	4	–	–	4	3	–	–	4	3
50–100	5	–	–	5	4	–	–	5	3
100–150	6	–	–	6	5	–	–	6	3

^a Horizon identification number.

configuration of the hypothetical pedon described above would be for UWMs with a maximum depth of 30 cm and for DWMs with a depth greater than 30 cm.

Additionally, for each database configuration (CD and DD), multiple upward and downward models were created in an attempt to identify any threshold depth for each PTF type. Previous work has suggested that individual PTFs should be created for horizons above and below the threshold depth of 30 cm due to the influence of soil management practices and plant root growth on ρ_b at 0–30 cm depth (Benites et al., 2007; Watson et al., 2000). For UWMs the maximum bottom depth of the horizon of interest was set at 20, 30, or 50 cm (UWM-20 for CD and DD, UWM-30 for CD and DD, and UWM-50 for CD and DD, respectively), with horizon(s) with an available ρ_b always being deeper than the horizons of interest. For DWMs the maximum top depth of the horizon of interest was set at 20, 30, or 50 cm (DWM-20 for CD and DD, DWM-30 for CD and DD, and DWM-50 for CD and DD, respectively), with horizon(s) with an available ρ_b always being shallower than the horizons of interest.

The combination of PTF type (UWM and DWM), data configuration (CD and DD), and data depth-split (20, 30, and 50 cm) led to a final number of 12 PTFs. The samples size (number of horizons) for each PTF type depth-split combination was different as there is a different availability of data with depth (Table 4).

2.5. Pedotransfer function validation

For all PTFs, 70% of the horizons were randomly selected for calibration and the remaining 30% used for validation. The prediction quality of the 12 PTFs was evaluated by the following parameters: prediction coefficient of determination (R^2), Eq. (1); root mean squared prediction error (RMSPE), Eq. (2); and mean prediction error (MPE), Eq. (3). These are defined as:

$$R^2 = \frac{[\text{cov}(\rho_{b,i}, \widehat{\rho}_{b,i})]^2}{\text{var}(\rho_{b,i}) * \text{var}(\widehat{\rho}_{b,i})} \quad (1)$$

$$\text{RMSPE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\widehat{\rho}_{b,i} - \rho_{b,i})^2} \quad (2)$$

$$\text{MPE} = \frac{1}{n} \sum_{i=1}^n (\widehat{\rho}_{b,i} - \rho_{b,i}) \quad (3)$$

where $\rho_{b,i}$ and $\widehat{\rho}_{b,i}$ are the observed and predicted ρ_b values, respectively; n is the number of observations; and var and cov are the variance and covariance, respectively. The R^2 measures the strength of the linear relationship between observed and predicted ρ_b values. The RMSPE measures the overall prediction error (accuracy) of the PTFs. The MPE

allows the evaluation of positive or negative bias of a PTF, indicating an average tendency for overestimation or underestimation, respectively (De Vos et al., 2005; Jalabert et al., 2010).

2.6. Assessing variable importance

To assess which subset of variables best explain ρ_b estimation, variable importance in the 12 PTFs was measured by setting the importance argument as “TRUE” in the randomForest function of R's randomForest package (R Development Core Team, 2011). Computations are carried out one tree at a time. The regression tree is constructed from the bootstrap sample. After tree construction, samples not selected in the bootstrap sample [out-of-bag (OOB) samples] are dropped down the tree and the mean squared error (MSE) of prediction is computed. Next, the MSE is computed again for the OOB samples after randomly permuting values/categories (changing positions) on the j th variable X_j while leaving the data on all other variables unchanged. By randomly permuting the predictor variable X_j , its original association with the response Y is broken. The percentage of increase in MSE is then averaged for the whole forest. If variable X_j is important, permuting its observed values/categories will reduce the ability to predict each OOB observation and, hence, increase the MSE when compared to the unaltered data (Izenman, 2008).

2.7. R codes

To facilitate transferability and adoption of the proposed PTFs, the platform codes used in R to develop the models are listed below:

```
> library(randomForest)
> formula <- as.formula(paste("rho_b ~", paste(names(calibration.set[,c(3:11)]), collapse="+")))
> model <- randomForest(formula, data= calibration.set, ntree=1500, importance=TRUE)
> prediction <- predict(model, newdata=validation.set)
> varImpPlot(model)
```

The first line is used to load the randomForest package. The second line defines the formula that establishes the relationship between the response variable (ρ_b of the horizon of interest) and the regressors. In this case example, the nine regressors were located at columns 3–11 of the dataset. In the third line, the PTF is fitted using the randomForest function and the formula defined above. In the fourth line, the ρ_b prediction is performed on the validation dataset using the predict function and the model fitted in the previous step. The fifth line returns the variable importance plot by using the varImpPlot function.

3. Results and discussion

3.1. Validation of the proposed pedotransfer functions

The validation performance of the 12 proposed PTFs is presented in Table 5. Upward models had R^2 , RMSPE, and MPE ranging at 0.50–0.59, 0.13–0.15 g cm^{−3}, and −0.013–−0.006 g cm^{−3}, respectively. The CD configuration resulted in slightly more accurate models than the DD configuration (Table 5). For instance, the RMSPE of UWM-20 derived from CD and DD configuration was 0.14 and 0.15 g cm^{−3}, respectively. The better performance of models derived from the CD configuration is due to the fact that horizons with an available ρ_b measurement are always adjacent to the horizons of interest (in this configuration), which is not always the case in the DD configuration. By always using adjacent horizons, the continuity of ρ_b in the soil profile is better exploited in the CD configuration, resulting in better (more accurate) ρ_b predictions. Using the k-nearest neighbor (k-NN) algorithm, Nemes et al. (2010) used sample depth for grouping reference samples before

Table 4
Descriptive statistics for soil bulk density (g cm^{−3}) in different dataset arrangements.

	N ^a	Mean	SD ^b	Minimum	Maximum
UWM-20 ^c	5423	1.34	0.22	0.23	2.04
UWM-30	6883	1.35	0.21	0.23	2.04
UWM-50	9003	1.37	0.20	0.23	2.04
DWM-20	13,634	1.44	0.19	0.44	2.41
DWM-30	11,805	1.45	0.20	0.44	2.33
DWM-50	9300	1.46	0.20	0.44	2.33

^a Sample size.

^b Standard deviation.

^c Upward models (UWMs) developed for maximum depths of 20, 30, and 50 cm (UWM-20, UWM-30, and UWM-50, respectively); downward models (DWMs) developed for depths below 20, 30, and 50 cm (DWM-20, DWM-30, and DWM-50, respectively).

Table 5

Validation results for pedotransfer functions (PTFs) developed for predicting soil bulk density.

PTF	R^2	RMSPE ^a	MPE ^b
		g cm^{-3}	
UWM-20 CD ^c	0.56	0.14	−0.011
UWM-30 CD	0.56	0.14	−0.010
UWM-50 CD	0.59	0.13	−0.006
UWM-20 DD	0.50	0.15	−0.013
UWM-30 DD	0.50	0.15	−0.010
UWM-50 DD	0.50	0.14	−0.006
DWM-20 CD	0.67	0.11	−0.009
DWM-30 CD	0.69	0.11	−0.006
DWM-50 CD	0.71	0.10	−0.005
DWM-20 DD	0.58	0.13	−0.007
DWM-30 DD	0.60	0.12	−0.008
DWM-50 DD	0.66	0.11	−0.008

^a Root mean squared prediction error.^b Mean prediction error.^c Upward models (UWMs) developed for maximum depths of 20, 30, and 50 cm (UWM-20, UWM-30, and UWM-50, respectively); downward modes (DWMs) developed for depths below 20, 30, and 50 cm (DWM-20, DWM-30, and DWM-50, respectively); continuous database (CD) configuration that mimics randomly missing bulk density values; discontinuous database (DD) configuration that mimics pattern-consistent missing bulk density values.

ρ_b estimation. In this approach, they attempted to optimize the match of the target sample depth to the reference sample depth for optimal model performance. Up to a certain extent, the close match between

depths could be considered similar to the CD configuration while the greater mismatch could be considered similar to the DD configuration. These authors observed, however, a disadvantage of restricting the reference samples within the range of 5 to 20 cm compared to the range of 25 to 60 cm. The disagreement between the two approaches is likely related to the incorporation of horizon continuity in the present study by always using adjacent horizons in the CD configuration.

For each database configuration (CD and DD), it was found that UWMs with a maximum horizon bottom depth of 50 cm (UWM-50) tended to perform better than those using maximum bottom depths of 20 and 30 cm (Table 5). However, all UWMs had satisfactory performance when compared to similar studies. Huang et al. (2003) reported that the majority of published studies, that have employed an acceptable error between 10 and 20% of the observed mean ρ_b for PTFs, to be useful. Additionally, De Vos et al. (2005) suggested that RMSPE should be less than 10% of the observed mean ρ_b for studies that require accurate calculations. In the present study, the RMSPE of all UWMs fell within the range of 10–20% suggested by Huang et al. (2003), but only the UWM derived from the CD configuration to a maximum horizon bottom depth of 50 cm (UWM-50 CD) was below the 10% limit as suggested by De Vos et al. (2005) (Tables 4 and 5).

Fig. 1 displays the performance of all UWMs, with models derived from CD configuration tending to have a tighter fit about the 1:1 relationship line (of a perfect prediction) than those derived from the DD configuration. Prediction values rotated around the 1:1 line;

Fig. 1. Performance of proposed pedotransfer functions developed to predict bulk density for near-surface horizons using information of the horizon of interest and a horizon below it. Upward models (UWMs) developed for maximum depths of 20, 30, and 50 cm (UWM-20, UWM-30, and UWM-50, respectively). Continuous database (CD) configuration mimics randomly missing bulk density values. Discontinuous database (DD) configuration mimics pattern-consistent missing bulk density values. The solid line represents the 1:1 relationship line.

overestimating low observed ρ_b values and underestimating high observed ρ_b values. This typical trend is credited to the different proportion of near-surface and sub-surface horizons in the dataset and to the fact that all PTF techniques tend to bias toward the grand mean of the response variable (De Vos et al., 2005; Nemes et al., 2010). All models had a slightly negative MPE (Table 5) indicating a slight overall underestimation of ρ_b . For both database configurations, UWMs at low ρ_b values were dispersed while, in general, predictions above 0.75 g cm^{-3} tended to cluster closer to the 1:1 line. Horizons with dispersed predictions were identified as A, Ap, and Bh horizons with SOC contents ranging between 13 and 284 g kg^{-1} . These horizons were further re-checked for bad ρ_b determination or bad data entry but no indication of bad data was detected. It is possible that the low number of horizons with ρ_b lower than 0.75 g cm^{-3} resulted in a sub-optimum learning process of models in this range and, hence, reduced prediction performance. Other non-parametric algorithms such as k-NN are credited not to be affected by data density (Nemes et al., 2010); however, the overall accuracy of the proposed PTFs overcomes this issue.

Downward models had R^2 , RMSPE, and MPE ranging between 0.58–0.71, $0.10\text{--}0.13 \text{ g cm}^{-3}$, and $-0.009\text{--}-0.005 \text{ g cm}^{-3}$, respectively (Table 5). Similar to the UWMs, deriving DWMs from CD configuration resulted in slightly more accurate models than deriving them from DD configuration, with the maximum horizon top depth of 50 cm (DWM-50) tending to be better than 20 and 30 cm for both database configurations (Table 5). The RMSPE of all DWMs represented less than 10% of the observed mean ρ_b (Tables 4 and 5); meaning that

all models would be adequate for studies requiring accurate calculations according to the criterion suggested by De Vos et al. (2005). Considering the threshold depth results for both upward and downward models, 50 cm was found to be the best depth to split the database between the data used for fitting UWMs (0–50 cm depth) and the data used for fitting DWMs ($\geq 50 \text{ cm}$ depth). This threshold depth is deeper than the 30 cm depth suggested by others (Benites et al., 2007; Watson et al., 2000), and is likely the result of the drastic decrease in root distribution for soil depths greater than 50 cm (Grant et al., 2012; Monti and Zatta, 2009). The accuracy of DWMs was slightly better than that of UWMs as the RMSPE ranged between 0.13–0.15 and $0.10\text{--}0.13 \text{ g cm}^{-3}$ for UWM and DWM, respectively (Table 5). The slightly better performance of DWMs is likely related to the fact that the majority of ρ_b values used for developing it were from sub-surface horizons located at depths below 30 cm. For instance, it would be necessary to use horizons at 20–30, 30–50, 50–100, and 100–150 cm depths to develop a DWM with the maximum horizon top depth of 20 cm (DWM-20) using the hypothetical pedon described in Table 3. On the other hand, a large number of ρ_b values used for developing UWMs were from near-surface horizons (e.g., above 20 cm depth) that are considerably affected by factors not included in the proposed PTFs (e.g., soil management, plant roots). In this case, the horizons at 0–5, 5–20, 20–30, and 30–50 cm depth would be needed to develop a UWM with a maximum horizon bottom depth of 50 cm (UWM-50) using the hypothetical pedon described in Table 3.

Fig. 2 displays the performance of all DWMs. As was observed with UWMs, DWMs derived from the CD configuration followed the 1:1

Fig. 2. Performance of proposed pedotransfer functions developed to predict bulk density for sub-surface horizons using information of the horizon of interest and a horizon above it. Downward modes (DWMs) developed for depths below 20, 30, and 50 cm (DWM-20, DWM-30, and DWM-50, respectively). Continuous database (CD) configuration mimics randomly missing bulk density values. Discontinuous database (DD) configuration mimics pattern-consistent missing bulk density values. The solid line represents the 1:1 relationship line.

Fig. 3. Variable importance measured by percentage increase in mean squared error (MSE) due to permutation of variables in pedotransfer function (PTFs). Upward model (UWM) developed for maximum depth of 50 cm (UWM-50). Downward mode (DWM) developed for depth below 50 cm (DWM-50). Continuous database (CD) configuration mimics randomly missing bulk density values. Discontinuous database (DD) configuration mimics pattern-consistent missing bulk density values. Horizon of interest (HOI) and horizon with available bulk density (HAD) in the soil profile.

relationship line slightly more closely than those derived from DD configuration. All models presented a negative MPE that indicates an overall underestimation of ρ_b (Table 5). Downward models did not have increased dispersion in the low range of ρ_b values as was shown for UWMs. It is likely that the absence of few and extreme low ρ_b values in the DWM datasets allowed for a better learning process in the low range of ρ_b values when fitting the models. On the other hand, except for DWM-50 DD, modeling downwards resulted in some dispersion in predicting high values of ρ_b . Horizons with $\rho_b \geq 2.0 \text{ g cm}^{-3}$ were identified as Bx, C, Cd, and Cr horizons. Again, a lack of representative samples in the high range of ρ_b values did not allow models to have the best possible learning process and, hence, resulted in sub-optimum accuracy. Tranter et al. (2007) set their lower and upper ρ_b limits at 0.7 and 1.8 g cm^{-3} , respectively, suggesting that samples outside this range would be dubious or exceptional. However, as stated in this study, extreme or dubious ρ_b values were re-checked for bad ρ_b determination or bad data entry and it was determined that they represent reliable measurements. Thus, these types of horizons may be present and in need of prediction for some databases.

The unique approach to take in consideration the information from a surrounding horizon (above or below and adjacent or not to the horizon of interest) gives the opportunity to accurately predict ρ_b for incomplete databases, assuming that some ρ_b measurements are available. The good performance of the proposed PTFs was obtained with the use of information that is commonly found in incomplete databases (e.g., horizon designation, textural class, and depth), without the need to have complicated data. Also using the USDA-NRCS National Soil Survey Center characterization database, Nemes et al. (2010) obtained a RMSPE of about 0.17 g cm^{-3} for their best modeling option. The range of RMSPE in present study was of $0.10\text{--}0.15 \text{ g cm}^{-3}$, indicating that the proposed approach offers an improvement in accuracy to predict ρ_b .

3.2. Variable importance

Based on the increase in MSE after permuting variable values, the most important variable in predicting ρ_b for all 12 PTFs was the ρ_b of horizons above or below the horizon of interest (Fig. 3). This is related to the continuity of ρ_b values in the soil profile making it possible to

interpolate/predict the ρ_b of a horizon of interest knowing the ρ_b of another horizon in the soil profile. In general, a higher increase in MSE was observed for the CD configuration than that for the DD configuration. For instance, permuting ρ_b values increased MSE by 336 and 270% for UWM-50 CD and UWM-50 DD, respectively, and by 463 and 339% for DWM-50 CD and DWM-50 DD, respectively (Fig. 3). This is likely from the fact that horizons with an available ρ_b were always adjacent to horizons of interest in the CD configuration, but not always in the DD configuration. Additionally, there is an apparent higher importance of ρ_b of another horizon (above or below and adjacent or not adjacent to the horizon of interest) in the soil profile for DWMs than for UWMs (Fig. 3). This is again related to a large number of ρ_b values in UWMs being from near-surface horizons (e.g., above 20 cm depth) that are considerably affected by factors not included in the proposed models (e.g., soil management, plant roots).

The textural class of the horizon of interest was, across all models, the second most important variable in predicting ρ_b , while textural class of the horizon with available ρ_b was, in general, the third most important variable (Fig. 3). Additionally, permuting textural class in the DD configuration resulted in a slightly greater increase in MSE than in the CD configuration due to the reduced importance of the ρ_b variable in the DD configuration, as described above. The majority of previous studies have not used textural classes as a variable for predicting ρ_b but used the contents of clay, silt, and/or sand, which determine textural classes (De Vos et al., 2005; Jalabert et al., 2010; Tranter et al., 2007). There is a wide range in the particle-size distribution for a textural class, making it suitable only for general classification purposes (Posadas et al., 2001). Still, textural class proved to be of great value in predicting ρ_b in the present study. Calhoun et al. (2001) used field textural class (tactile determinations) to predict ρ_b in their proposed “Field Model” and found it to be the second most important variable, explaining 6% of the variability. Pachepsky and Rawls (1999) also found that grouping soil by textural class improved the accuracy of PTFs in predicting volumetric soil water contents. The importance of soil texture in predicting ρ_b relates to the general tendency that fine-

textured surface soils have lower ρ_b than sandy soils due to a higher total pore space (in clayey soils than in sandy soils), which, for the present study, is illustrated in Fig. 4.

In general, horizon designation of the horizon of interest and of the other horizons in the soil profile ranked as the fourth and fifth most important variables in predicting ρ_b , respectively. Using site and morphology data, Calhoun et al. (2001) found horizon designation to be the most important variable in predicting ρ_b , accounting for 49% of the variability. Horizons are differentiated to reflect differences in soil properties that may or may not be the result of soil formation (Soil Survey Division, 1993). For instance, A horizons are generally characterized by i) being mineral horizons that exhibit obliteration of all or much of the original rock structure, ii) presenting accumulation of humified organic matter mixed with the mineral fraction, and iii) presenting properties resulting from cultivation, pasturing, or similar kind of disturbance. These characteristics tend to result in A horizons having a lower ρ_b than deeper horizons that preserve more of the original parent material structure and have lesser amounts of humified organic matter and lesser disturbance. Fig. 5 illustrates the relationship between the horizon designation categories and the measured ρ_b data used in the present study. Among mineral horizons, A horizons had the lowest median ρ_b value compared to the other horizons except Bhs horizons, likely due to illuvial accumulation of organic matter in the latter category.

Depth of the horizon of interest, depth of the horizon with available ρ_b , thickness of the horizon of interest, and thickness of the horizon with available ρ_b occupied the remaining positions of variable importance, with an overall greater importance of depth over thickness. The lesser importance of depth in predicting ρ_b in the present study is related to the large range of ρ_b at the various depths and the weak relationship between these soil properties (Fig. 4). Our results agree with those of De Vos et al. (2005) and Heuscher et al. (2005) who also did not find depth to be an important (significant) factor in predicting ρ_b . Among continuous variables in the dataset, horizon thickness presented the weakest relationship with observed ρ_b values (Fig. 4).

Fig. 4. Scatter plots of observed bulk density vs. continuous variables.

Fig. 5. Box-and-whisker plot of observed bulk density values for the 20 horizon designation categories. The horizontal line represents the median bulk density for each horizon designation category. The bottom and top of the box represent the 25th and 75th percentiles, respectively. Whiskers extend 1.5 times the interquartile range. Sample data fall outside the whiskers are referred to as outliers. The medians of two box-and-whiskers plots are significantly different at approximately the 0.95 confidence level if the corresponding notches do not overlap.

4. Conclusions

Random forest models were successfully used to create novel PTFs that incorporated information, mainly ρ_b , from other horizons in the soil profile to predict missing ρ_b values. These relatively accurate PTFs can be used to help researchers and surveyors fill gaps in their databases without complicated data acquisitions.

In addition to the ρ_b of a specific horizon (above or below and adjacent or not adjacent to the horizon of interest), the variables used for developing the PTFs were textural class and basic pedon description information (horizon designation, and top and bottom horizon depths) that are commonly found in databases. A threshold depth of 50 cm was found to give the best upward and downward predicting models for two database configurations (continuous and discontinuous), but with continuous database configuration giving slightly better PTFs. Thus, the best upward and downward models resulted in RMSPE of 0.13 and 0.10 g cm⁻³, respectively. Available ρ_b within the soil profile was the most important variable in predicting ρ_b , reaffirming the applicability of the proposed PTFs for databases not missing all ρ_b measurements. The availability of textural class and basic pedon description information from the horizon of interest (the horizon for which ρ_b is being predicted), and ρ_b , textural class, and basic pedon description information from horizons above or below and adjacent or not adjacent to the horizon of interest allowed for the development of reasonably accurate PTFs (RMSPE ranging from 0.10 to 0.15 g cm⁻³) without the need of SOC, a variable often used to predict ρ_b .

Acknowledgments

This research was sponsored by the USDA-NRCS. The authors would like to thank all staff of the National Soil Survey Center for sampling, analyzing, and managing the data used in this research.

References

- Adams, W.A., 1973. The effect of organic matter on the bulk density and true densities of some uncultivated podzolic soils. *Journal of Soil Science* 24, 10–17.

- Aertsen, W., Kint, V., van Orshoven, J., Özkan, K., Muys, B., 2010. Comparison and ranking of different modeling techniques for prediction of site index in Mediterranean mountains forest. *Ecological Modelling* 221, 1119–1130.
- Benites, V.M., Machado, P.L.O.A., Fidalgo, E.C.C., Coelho, M.R., Madari, B.E., 2007. Pedotransfer functions for estimating soil bulk density from exiting soil survey reports in Brazil. *Geoderma* 139, 90–97.
- Breiman, L., 1996. Bagging predictors. *Machine Learning* 24, 123–140.
- Breiman, L., 2001. Random forests. *Machine Learning* 45, 5–32.
- Calhoun, F.G., Smeck, N.E., Slater, B.L., Bigham, J.M., Hall, G.F., 2001. Predicting bulk density of Ohio soils from morphology, genetic principles, and laboratory characterization data. *Soil Science Society of America Journal* 65, 811–819.
- Chan, K.Y., Kwong, C.K., Dillon, T.S., Tsim, Y.C., 2011. Reducing overfitting in manufacturing process modeling using a backward elimination based genetic programming. *Applied Soft Computing* 11, 1648–1656.
- Dam, R.F., Mehdi, B.B., Burgess, M.S.E., Madramootoo, C.A., Mehuys, G.R., Callum, I.R., 2005. Soil bulk density and crop yield under eleven consecutive years of corn with different tillage and residue practices in a sandy loam soil in central Canada. *Soil and Tillage Research* 84, 41–53.
- De Vos, B., Meirvenne, M.V., Quataert, P., Deckers, J., Muys, B., 2005. Predictive quality of pedotransfer functions for estimating bulk density of forest soils. *Soil Science Society of America Journal* 69, 500–510.
- Ellert, B.H., Bettany, J.R., 1995. Calculation of organic matter and nutrient stored in soils under constrain management regimes. *Canadian Journal of Soil Science* 75, 529–538.
- Foss, J.E., Wright, W.R., Coles, R.H., 1975. Testing and accuracy of field textures. *Soil Science Society of America Proceedings* 39, 800–802.
- Gál, A., Vyn, T.J., Michéli, E., Kladičko, E.J., McFee, W.W., 2007. Soil carbon and nitrogen accumulation with long-term no-till versus moldboard plowing overestimated with tilled-zone sampling depths. *Soil and Tillage Research* 96, 42–51.
- Grant, J.C., Nichols, J.D., Yao, R.L., Smith, R.G.B., Brennan, P.D., Vanclay, J.K., 2012. Depth distribution of roots of *Eucalyptus dunnii* and *Corymbia citriodora* subsp. *variegata* in different soil conditions. *Forest Ecology and Management* 269, 249–258.
- Heuscher, S.A., Brandt, C.C., Jardine, P.M., 2005. Using soil physical and chemical properties to estimate bulk density. *Soil Science Society of America Journal* 69, 51–56.
- Ho, T.K., 1998. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 832–844.
- Huang, S., Yang, Y., Wang, Y., 2003. A critical look at procedures for validating growth and yield models. In: Amaro, A., Reed, D., Soares, P. (Eds.), *Modeling Forest Systems*. CAB International, Wallingford, UK, pp. 271–293.
- Izenman, A.J., 2008. Committee machines. In: Izenman, A.J. (Ed.), *Modern Multivariate Statistical Techniques*. Springer, New York, NY, pp. 505–549.
- Jalabert, S.S.M., Martin, M.P., Renaud, J.P., Boulonne, L., Jolivet, C., Montanarella, L., Arruays, D., 2010. Estimating forest soil bulk density using boosted regression modeling. *Soil Use and Management* 26, 516–528.
- Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. The newsletter of the R project, 2 18–21.
- Luellen, J.K., Shadish, W.R., Clark, M.H., 2005. Propensity scores: An introduction and experimental test. *Evaluation Review* 29, 530–558.

- Monti, A., Zatta, A., 2009. Root distribution and soil moisture retrieval in perennial and annual energy crops in Northern Italy. *Agriculture, Ecosystems and Environment* 132, 252–259.
- Nemes, A., Quebedeaux, B., Timlin, D.J., 2010. Ensemble approach to provide uncertainty estimates of soil bulk density. *Soil Science Society of America Journal* 74, 1938–1945.
- Pachepsky, Y.A., Rawls, W.J., 1999. Accuracy and reliability of pedotransfer functions as affected by grouping soils. *Soil Science Society of America Journal* 63, 1748–1757.
- Posadas, A.N.D., Giménez, D., Bittelli, M., Vaz, C.M.P., Flury, M., 2001. Multifractal characterization of soil particle size distribution. *Soil Science Society of America Journal* 65, 1361–1367.
- Post, D.F., Huete, A.R., Pease, D.S., 1986. A comparison of soil scientist estimations and laboratory determinations of some Arizona soil properties. *Australian Journal of Soil and Water Conservation* 41, 421–424.
- R Development Core Team, 2011. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rawls, W.J., 1983. Estimating soil bulk density from particle size analysis and organic matter content. *Soil Science* 135, 123–135.
- Reichert, J.M., Suzuki, L.E.A.S., Reinert, D.J., Horn, R., Håkansson, I., 2009. Reference bulk density and critical degree-of-compactness for no-till crop production in subtropical highly weathered soils. *Soil and Tillage Research* 102, 242–254.
- Schoeneberger, P.J., Wysocki, D.A., Benham, E.C., Broderson, W.D., 2002. *Field Book for Describing and Sampling Soils, Version 2.0*. NRCS-National Soil Survey Center, Lincoln, NE.
- Soil Survey Division, 1993. *Soil survey manual. Handbook 18*. USDA-Natural Resources Conservation Center, Washington, DC.
- Soil Survey Staff, 2004. *Soil survey laboratory methods manual. Soil Survey Investigations Report. 42, Version 4.0*. NRCS, Washington, DC.
- Tranter, G., Minasny, B., Mcbratney, A.B., Murphy, B., McKenzie, N.J., Grundy, M., Brough, D., 2007. Building and testing conceptual and empirical models for predicting soil bulk density. *Soil Use and Management* 23, 437–443.
- Verhulst, N., Govaerts, B., Verachtert, E., Castellanos-Navarrete, A., Mezzalama, M., Wall, P.C., Chocobar, A., Deckers, J., Sayre, K.D., 2010. Conservation agriculture, improving soil quality for sustainable production systems? In: Lal, R., Stewart, B.A. (Eds.), *Food Security and Soil Quality*. CRC Press, Boca Raton, FL, pp. 137–208.
- Watson, R.T., Noble, I.R., Bolin, B., Ravindranath, N.H., Verardo, D.J., Dokken, D.J., 2000. *Land Use, Land-Use Change, and Forest*. Cambridge University Press, Cambridge, UK (377 pp.).
- Zacharias, S., Wessolek, G., 2007. Excluding organic matter content from pedotransfer predictors of soil water retention. *Soil Science Society of America Journal* 71, 43–50.