Brazilian Soil Bulk Density Prediction Based on a Committee of Neural Regressors

Diego B. Haddad*, Laura S. de Assis*, Luis Tarrataca*, Andrea da S. Gomes†, Marcos Bacis Ceddia†,
Rosane F. Oliveira†, Jurair R. de P. Junior* and Diego N. Brandão*

*Federal Center of Technological Education of Rio de Janeiro - CEFET/RJ.
{*diego.haddad, laura.assis, luis.tarrataca, jurair.junior, diego.brandao*}@*cefet-rj.br*

†Federal Rural University of Rio de Janeiro - UFRRJ.
andrea_zooufc@yahoo.com.br, {*ceddia, rosaneol*}@*ufrrj.br*

***Abstract*—Computer models have been an important tool to determine soil bulk density. This soil property is fundamental to estimate soil carbon reserves and consequently to understand the global carbon cycle. The estimation of soil bulk density is not a trivial task since it demands an intensive and often impractical work. The purpose of this paper is to evaluate the performance of a pedotransfer function against an Artificial Neural Networks to estimate soil bulk density for soils at Brazilian biomes. The first one consists of a linear model composed of a Least Square method. The latter employs a robust committee of multilayer perceptron networks and a model selection procedure based on k-fold cross-validation. The data are composed of 3404 soil layers distributed in different Brazilian regions and with different uses. The proposed non-linear regressor presents higher precision when compared to the linear model, and requires less information to do so. Additionally, the developed solution brings to light the assumed relationship between soil bulk density and some soil chemical properties.**

***Index Terms*—Soil properties, Soil bulk density, Pedotransfer functions, Multilayer perceptron artificial neural network.**

I. Introduction

Increasingly, there is a demand for information to assist sustainable agriculture and improve the land quality. According to Budiman *et al.* [1] the usefulness of soil survey is not limited to producing data on inventories and geographical distribution of soil classes, but also to provide the quantitative spatial distribution of soil properties (such as clay content, soil density, saturated hydraulic conductivity, and available water capacity).

Among these properties, one of the most important is soil bulk density. This property is a physical characteristic that allows the determination of: soil hydraulic potential, root growth and particularly the specification of the amount of Soil Organic Carbon (SOC) stock [2]–[4], so that it can act as a source of atmospheric CO_2 and influence the draining of greenhouse gases [5], [6].

The determination of soil bulk density (D_s) is a hard task that requires time-consuming laboratory analysis [2]. Another difficulty is related to the fact that D_s presents high spatial and temporal variability [7]. Accordingly, in general, D_s is not included in soils databases.

Since D_s is difficult to measure, inferential modeling provides an interesting alternative that surmises the variable of interest using other more easily measurable ones [8]. Pedotransfer functions (PTFs), based on easily measured soil attributes, show strong potential to replace D_s measurements when their direct measurement is not feasible [3].

PTFs are predictive mathematical functions of certain soil properties, including soil bulk density, which allows estimating D_s from other soil features more easily measured and routinely obtained at lower costs. The PTFs fulfill the demands for data normally available in soil surveys and databases.

According to Qinna *et al.* [4], PTFs models can be divided into four types: (1) physical-conceptual modeling approaches; (2) linear or non-linear regression models; (3) multiple regression methods and (4) advanced models, such as artificial neural networks (ANNs). More details on the development and use of PTFs can be found at [9] and [10].

Some studies tested the performance of available PTFs and noticed that these functions are somewhat inaccurate when applied to different environments [11] and [2]. Physical models need a detailed database and *a priori* knowledge about how each soil property affects other properties [6]. Qinna et al. [4] showed that linear models have low prediction capabilities in some problems and the non-linear models were able to significantly enhance the accuracy of the prediction in comparison to linear models. Advanced models, such as ANNs, demonstrate a great power of generalization, which is an attractive characteristic to determine D_s in the different Brazilian biomes.

This paper proposes to evaluate the performance of linear and non-linear PTFs to estimate D_s for soils in the Brazilian territory. Linear and nonlinear regressions are compared. The latter employs a robust committee of multilayer perceptrons networks and a model selection procedure based on k-fold cross-validation.

The paper is organized as follows. Related works are described in Section II. The notation employed in this paper is presented in Section III. Section IV describes the Least Squares Regression model. The proposed solution based on Multilayer Perceptron ANN is presented in Section V. Section VI gives the computational experiments and the analysis of the results. Conclusions and final remarks follow in Section VII.

II. Related Works

PTFs are an important technique to determine soil properties. In this context, several works were developed using PTFs to determine D_{s} [4], [12]–[15]. In particular, concerning the Brazilian case, the main works were developed by [2], [3], [6] [16], [17] and [11].

Benites *et al.* [3] proposed in their paper a simplified regression model to predict D_{s} from available soil properties present in most biomes. The authors used two data sets: the first one constructed from the Soil Archives of Embrapa Solos in Rio de Janeiro, Brazil and the second being an independent soil dataset organized from the International Soil Classification Workshops. The proposed model was compared with three existing estimating models present in the literature. The results showed that the proposed simplified regression model was less biased and more accurate compared to the three existing regression equations.

Barros *et al.* [16] presented stepwise regression models to estimate soil bulk density using data on soil carbon and clay content and pH in water. The results were compared with those present in the literature and showed that local-based regressions are the most accurate for estimating D_{s} upland forests in the Manaus region.

Boschi *et al.* [2] evaluate the predictive capability of 25 PTFs available in the literature to estimate D_{s} in different regions of Brazil with different metrics. The results showed that when the observed and estimated bulk density values are compared, the best results were found with BEN-C and M&J-B models.

Serqueira *et al.* [17] investigated the performance of new PTFs in predicting soil bulk density. The PTFs were created considering the direction of prediction in the soil profile (upward and downward) using a tree-based algorithm for developing the PTFs. The authors showed that the proposed PTFs are reasonably accurate and have the potential to help researchers and other users to fill gaps in their database without complicated data acquisition.

Gomes *et al.* [11] developed one PTF to estimate Ds in soils of the Brazilian Central Amazon region. Their model outperformed the other knowledge literature models. Moreira *et al.* [6] proposed a Least-Squares ANN to estimate D_{s} for soils in the Amazon Forest area. Although the LS regression is not robust against perturbations on data, the method presented enhanced precision when compared with results obtained using PTFs based on some works present in the literature that use the same database.

III. Notation

The notation employed in this paper is defined as follows. The input data points are concatenated in the vectors $\{\boldsymbol{x}_i\}_{i=1}^{n}$, where $\boldsymbol{x}_i \in \mathbb{R}^{N_0\times 1}$. The target/output variable is defined by $\{d_i\}_{i=1}^{n}$. The estimated target value is defined as $\{y_i\}_{i=1}^{n}$. The number of training and test samples is given by n_{train} and n_{test}, respectively. The operator $(\cdot)^{\text{T}}$ denotes transposition.

IV. Least-Squares Regression

Least squares (LS) regression is a technique widely employed in the machine learning area, due to its simple formulation, compact form, and efficient closed-form solution [18]. Such features give to LS regression a fundamental status in data processing and classification [19]. In short, LS prediction aims to estimate the parameters of a linear model in order to best fit the observed data, minimizing the sum of squared residual errors [20], [21]. A popular point of view regards LS regression as a special case of the more general method of maximum likelihood [22]. Let $\boldsymbol{X}_{\text{train}} \in \mathbb{R}^{n_{\text{train}}\times N_0}$ be the data matrix defined as:

$$\boldsymbol{X}_{\text{train}} \triangleq \left[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n_{\text{train}}}\right]^{\text{T}}. \tag{1}$$

The purpose of LS prediction is to learn a regression vector $\boldsymbol{w} \in \mathbb{R}^{N_0\times 1}$ and an offset $b \in \mathbb{R}$ in order to express approximately the target vector $\boldsymbol{d}_{\text{train}} \in \mathbb{R}^{n_{\text{train}}\times 1}$ as:

$$\boldsymbol{d}_{\text{train}} \approx \boldsymbol{X}_{\text{train}}\boldsymbol{w} + \boldsymbol{e}_n b, \tag{2}$$

where $\boldsymbol{e}_n \in \mathbb{R}^{n_{\text{train}}\times 1}$ is a vector whose elements are equal to one [23].

Although the LS regression is a powerful method provided with closed-form solution [24], it fails when the input-output relationship is nonlinear, there are several outliers and/or the measurement noise is not Gaussian [25]. The former issue is caused by the misleading assumption of linear relationships modeling the potential association between the dependent variable and each independent variable [26]. In order to overcome such constraint, ANNs could be employed. The neural architecture employed in this paper is described in the following section.

V. Multilayer Perceptron

A multilayer perceptron (MLP) ANN is a bioinspired machine learning approach widely employed in a variety of fields, such as gait recognition [27], long-term forecast of electricity demand [28], obesity prediction [29], drought forecasting [30], just to mention a few. An ANN is able to model complex nonlinear relations between the input data and the quantity one is interested in estimating [31]. The MLP architecture includes one or more hidden layers, one output, and one input layer. Each ANN layer presents multiple neurons, connected to other neurons through synaptic weights (see Figure 1). The state of each neuron is evaluated by a multiplication and accumulation procedure, which performs a weighted summation of the outputs of neurons from the previous layer using synaptic weights [32].

Let's define a shallow MLP as a network that presents three layers: the input one (layer 0), the hidden one (layer 1) and the output one (layer 2). The excitation vector $\boldsymbol{z}^{(1)}$ of the hidden layer is defined by:

$$\boldsymbol{z}^{(1)} \triangleq \boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}, \tag{3}$$

where $\boldsymbol{x} \in \mathbb{R}^{N_0\times 1}$ is the feature (or observation) vector, $\boldsymbol{W}^{(i)} \in \mathbb{R}^{N_i\times N_{i-1}}$ is the weight matrix, $\boldsymbol{b}^{(i)} \in \mathbb{R}^{N_i\times 1}$ is the

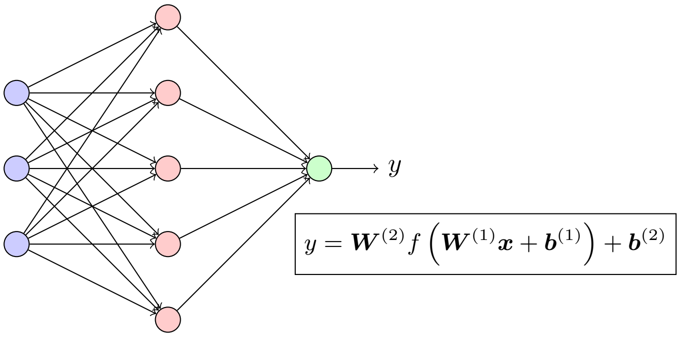

y
$y = \boldsymbol{W}^{(2)} f\left(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}\right) + \boldsymbol{b}^{(2)}$

Fig. 1. A functional block diagram of a shallow neural network with 3 input data (in blue), 5 hidden neurons (in red) and one output (in green). Typically, the output layer is linear in regression tasks, such as the one addressed in this paper.

ith-layer bias vector, and N_i is the number of neurons at the ith layer. The activation vector $\boldsymbol{v}^{(1)} \in \mathbb{R}^{N_1 \times 1}$ of the hidden layer can be evaluated from:

$$\boldsymbol{v}^{(1)} = f\left(\boldsymbol{z}^{(1)}\right), \tag{4}$$

where $f(\cdot) : \mathbb{R}^{N_1 \times 1} \rightarrow \mathbb{R}^{N_1 \times 1}$ is the activation function applied in an element-wise way. It must be stressed that the activation function should be nonlinear, otherwise the MLP degenerates into a LS prediction [33]. Some popular choices for $f(\cdot)$ are the hyperbolic tangent function, the sigmoid function, and the rectified linear unit (ReLU) function [34].

For regression problems, a popular empirical training criterion for the model parameters (i.e., weighting matrices $\boldsymbol{W}^{(i)}$ and bias vectors $\boldsymbol{b}^{(i)}$) is the mean square error (MSE). Such criterion is typically minimized by means of classical back-propagation algorithm (which can be derived from the chain rule for gradient computation) and its variants [35]. Two different phases (the forward - see Alg. 1 - and the backward one) compose the classical back-propagation algorithm[1] (see Alg. 2), which is based on first-order gradient information [36].

Algorithm 1 Shallow ANN Forward Computation

```
procedure ForwardComputation($\boldsymbol{X}$)
                        ▷ Each column of $\boldsymbol{X}$ is a feature vector
    $\boldsymbol{Z}^{(1)} \leftarrow \boldsymbol{W}^{(1)}\boldsymbol{X} + \boldsymbol{B}^{(1)}$
                        ▷ Each column of $\boldsymbol{B}^{(i)}$ is $\boldsymbol{b}^{(i)}$
    $\boldsymbol{V}^{(1)} \leftarrow f\left(\boldsymbol{Z}^{(1)}\right)$
            ▷ The $i$th column of $\boldsymbol{V}^{(1)}$ is the activation vector
            associated to the $i$th observation vector
    $\boldsymbol{Y} \leftarrow \boldsymbol{W}^{(2)}\boldsymbol{V}^{(1)} + \boldsymbol{B}^{(2)}$
    Return $\boldsymbol{Y}$
            ▷ The output vector related to the $i$th observation
            vector is the $i$th column of $\boldsymbol{Y}$
end procedure
```

Algorithm 2 Shallow ANN Backpropagation Batch Algorithm

```
procedure Backpropagation($\boldsymbol{X}, \boldsymbol{D}$)
                        ▷ Each column of $\boldsymbol{D}$ consists of target values
    $k \leftarrow 0$
    while stopping criterion not met do
        Call ForwardComputation($\boldsymbol{X}$)
        $\boldsymbol{G}_k^{(2)} \leftarrow \boldsymbol{D} - \boldsymbol{Y}$
        $\nabla_{\boldsymbol{W}_k^{(2)}} \mathcal{J}_{\text{MSE}} \leftarrow \frac{1}{\#\mathcal{T}} \sum_{m=1}^{\#\mathcal{T}} (\boldsymbol{d}_m - \boldsymbol{y}_m) \left(\boldsymbol{v}_m^{(1)}\right)^{\mathrm{T}}$
                        ▷ $\#\mathcal{T}$ is the cardinality of the training set
        $\nabla_{\boldsymbol{b}_k^{(2)}} \mathcal{J}_{\text{MSE}} \leftarrow \frac{1}{\#\mathcal{T}} \sum_{m=1}^{\#\mathcal{T}} (\boldsymbol{d}_m - \boldsymbol{y}_m)$
        $\boldsymbol{W}_{k+1}^{(2)} \leftarrow \boldsymbol{W}_k^{(2)} + \beta \nabla_{\boldsymbol{W}_k^{(2)}} \mathcal{J}_{\text{MSE}}$
                        ▷ $\beta$ is the adjustable step-size
        $\boldsymbol{b}_{k+1}^{(2)} \leftarrow \boldsymbol{b}_k^{(2)} + \beta \nabla_{\boldsymbol{b}_k^{(2)}} \mathcal{J}_{\text{MSE}}$
        $\boldsymbol{E}_k^{(1)} \leftarrow \left(\boldsymbol{W}_k^{(2)}\right)^T \boldsymbol{G}_k^{(2)}$
        $\boldsymbol{G}_k^{(1)} \leftarrow f'\left(\boldsymbol{Z}^{(1)}\right) \odot \boldsymbol{E}_k^{(1)}$
                        ▷ $f'(\cdot)$ is the derivative of $f(\cdot)$
                        ▷ $\odot$ is the element-wise product
        $\nabla_{\boldsymbol{W}_k^{(1)}} \mathcal{J}_{\text{MSE}} \leftarrow \frac{1}{\#\mathcal{T}} \sum_{m=1}^{\#\mathcal{T}} \boldsymbol{g}_{k,m}^{(1)} \boldsymbol{x}_m^T$
                        ▷ $\boldsymbol{x}_m$ is the $m$th observation vector
                        ▷ $\boldsymbol{g}_{k,m}^{(1)}$ is the $m$th column of $\boldsymbol{G}_k^{(1)}$
        $\nabla_{\boldsymbol{b}_k^{(1)}} \mathcal{J}_{\text{MSE}} \leftarrow \frac{1}{\#\mathcal{T}} \sum_{m=1}^{\#\mathcal{T}} \boldsymbol{g}_{k,m}^{(1)}$
        $\boldsymbol{W}_{k+1}^{(1)} \leftarrow \boldsymbol{W}_k^{(1)} + \beta \nabla_{\boldsymbol{W}_k^{(1)}} \mathcal{J}_{\text{MSE}}$
        $\boldsymbol{b}_{k+1}^{(1)} \leftarrow \boldsymbol{b}_k^{(1)} + \beta \nabla_{\boldsymbol{b}_k^{(1)}} \mathcal{J}_{\text{MSE}}$
        $k \leftarrow k + 1$
    end while
end procedure
```

A. *Model Selection Criterion*

Since a shallow MLP is chosen for regression purposes, determining the number of neurons situated in the hidden layer is a crucial question for its ability to generalize on future data outside the training set. Such a problem is intrinsically related to the bias/variance dilemma [37] and to the selection of the best fitting model from a set of candidate models [8]. The usage of an excessive number of basis functions will cause over-fitting, and it is well known that the addition of more hidden neurons is equivalent to adding more basis functions in function approximation settings [38]. In short, one must balance between the complexity of the model and performance of fit in order to obtain a good generalization [39].

The interesting geometric interpretation described in [40] provides several useful guidelines for the problem of determining the number of hidden neurons, but such method can be only employed to problems with input data dimensionality inferior to three. Introducing noise to the training vectors may improve the generalization capability of a network, but this approach was only demonstrated with preset network parameters [38], [41].

Cross-validation is a robust statistical technique for estimating the generalization error (or, equivalently, the true risk function [42]), the most important operational performance of

[1] The algorithms descriptions focus on regression tasks.

a trained network [43]. In this paper, k-fold cross-validation is employed for model selection (i.e., determination of the number of hidden neurons) purposes [44].

B. Committee Machine

Heuristic and simple ANN design procedures do not exploit the full potential of neural networks [45]. A committee machine is able to reduce the performance degradation that may occur due to the dependency of a specific initialization of an ANN (which may lead it to a suboptimal local minimum). This idea can be traced back to 1965 and can be classified into two major categories depending on whether the final output involves the input signal or not [46]. In this paper, a static committee machine for the soil density prediction is proposed. The algorithm we focus on is classified as static because it relies on an average-type learning mechanism to do the integration [46]. The intuition behind the approach consists of recognizing that a better regressor is formed by combining multiple weaker regressors and such a combination is the subject of ongoing intense research [47]. Although an average or convex combination of the individual estimates of the regressors can be employed [48], [49], the robustness and simplicity provided by the median statistic have proven to excel in robustness against outliers [50]. Accordingly, in this paper the median of the estimates of P concurrent ANNs will be employed as the global ensemble estimate[2] - see Figure 2.

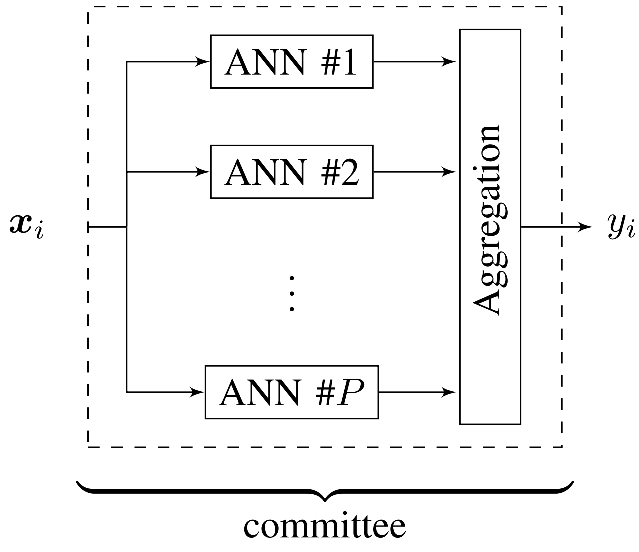


Fig. 2. Block diagram of the proposed committee machine. Note that the aggregation step consists of the median operator.

VI. Results

The data collection used in this paper consists of 3404 soil layers distributed in different Brazilian regions and with different uses (pasture, native vegetation, etc). The data were obtained with Soil Department of UFRRJ. The data presented

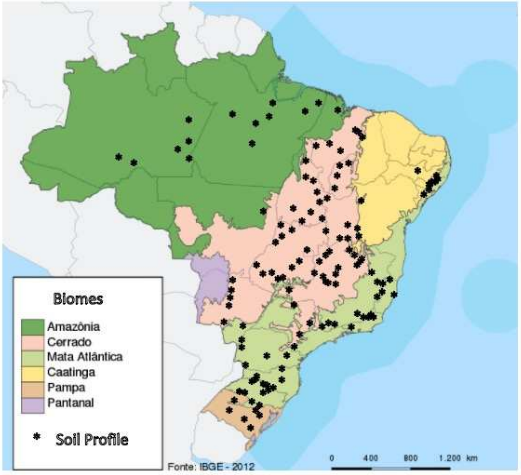


Fig. 3. Map with the location of points sampled. Adapted from [2], [3] and [6].

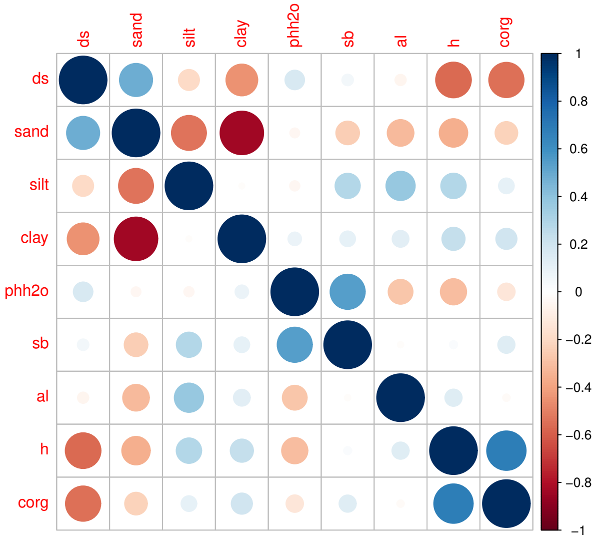


Fig. 4. Correlation matrix of the data.

soil properties obtained from a soil survey conducted in the Oil Province of Urucu River [6] and from the Soil Archives of Embrapa Solos in Rio de Janeiro [2], [3]. Figure 3 shows data distribution in the Brazilian territory.

Figure 4 shows the correlation matrix of the collected data. This figure reveals that there is a significant correlation (absolute value of approximately 0.5) between:

- ds and {sand, clay, h, corg}
- sand and {silt, clay}
- phh2o and {sb}
- h and {corg}

The LS solution obtained the following PTF:

[2] Note that the number of hidden neurons of each neural regressor will be chosen from the k-fold cross-validation technique.

$$
\begin{aligned}
y_i =& 1.0959 + 0.0005x_{1,i} + 0.0002x_{2,i} + 0.0001x_{3,i} \\
&+ 4.2895 \times 10^{-5}x_{4,i} + 0.0007x_{5,i} + 0.0003x_{6,i} \\
&- 0.0013x_{7,i} - 0.0095x_{8,i},
\end{aligned}
$$

where the physical meaning of input data $x_j(i)$ is described in Table I, where "Basic Cations" consist of Ca^{2+}, Mg^{2+}, K^{+} and Na^{+}.

TABLE I
INPUT DATA SOIL PROPERTIES.

Variable	Specification	Description
x_1	Sand	Comprising sand $(2.00 - 0.05$ mm)
x_2	Silt	Soil particle silt $(0.05 - 0.002$ mm)
x_3	Clay	Soil particle clay $(< 0.002$ mm)
x_4	pH	Chemical attribute pH (Water)
x_5	Sb	Sum of Basic Cations (cmol$_c$ dm^{-3})
x_6	Al^{3+}	Chemical attribute Aluminum cation (cmol$_c$ dm^{-3})
x_7	H^{+}	Chemical attribute Hydrogen cation (cmol$_c$ dm^{-3})
x_8	SOC	Soil Organic Carbon (g.kg^{-1})

The number of hidden neurons (chosen by 10-fold cross-validation procedure aiming at minimizing the mean square training error) is set to 6. This number was selected by inspection of the minimum mean-square errors in the training set. We chose $P = 7$ ANNs in parallel because of such a value was the most stable against initialization issues.

Each ANN is initialized with different weight values at random and trained with the LM (Levenberg-Marquardt) batch method with momentum [51], [52] and 200 epochs[3]. The hidden layer employs hyperbolic tangent activation functions and the output layer is a linear one.

The ith predicted value of one specific model is denoted by y_i. To assess the performance of the proposed regressors, the mean-squared prediction error (MSE, Equation (5)), the mean absolute error (MAE, Equation (6)) and the mean absolute percentage error (MAPE, Equation (7)) are employed. For a test set with target values d_i (for $i \in \{1, \cdots, n_{\text{test}}\}$), the definitions of these metrics are the following:

$$
\text{MSE} \triangleq \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (y_i - d_i)^2, \tag{5}
$$

$$
\text{MAE} \triangleq \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} |y_i - d_i|, \tag{6}
$$

$$
\text{MAPE} \triangleq \frac{100\%}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left| \frac{y_i - d_i}{y_i} \right|. \tag{7}
$$

Table II presents the MSE, MAE and MAPE results for the LS, LS′ and committee regressors. We distinguish between LS and LS′ where the former encompasses all the available variables described in the correlation matrix whereas the latter dispenses with the two least correlated variables with D_{s}, namely Sb and Al. The rationale being that the acquisition of fewer variables translates to less effort not only in terms of measurement but also regarding computational complexity and overfitting issues. The committee also employs the same set of variables used by LS′. In this way, both the LS′ and committee methods are in an equal standing.

Notice that the LS approach behaves best when all the variables are used. However, even when using all variables the committee still outperforms the LS approach. If we consider LS′, *i.e.*, when on an equal footing in terms of the variable set used, the committee also presents better performance. This is relevant since we are obtaining better results and at the same time requiring less information to do so.

It is also important to mention that there exists an additional metric, respectively the mean prediction error MPE [6]. The MPE metric enables the evaluation of an average tendency for underestimation (negative value) or overestimation (positive values), which indicates the signal of the predictor bias. The committee and LS solutions present very similar MPEs (-0.0144 and -0.014, respectively). Such values are slightly outperformed by the LS′ method (0.0133).

When the MPE metric is calculated for these methods the LS approach slightly outperforms the LS′ and committee methods.

TABLE II
EVALUATION OF THE LS REGRESSORS AND THE COMMITTEE.

Metric	LS	LS′	Committee
MSE	0.0218	0.0230	0.0197
MAE	0.1199	0.1239	0.1131
MAPE	9.48%	9.81%	8.95%

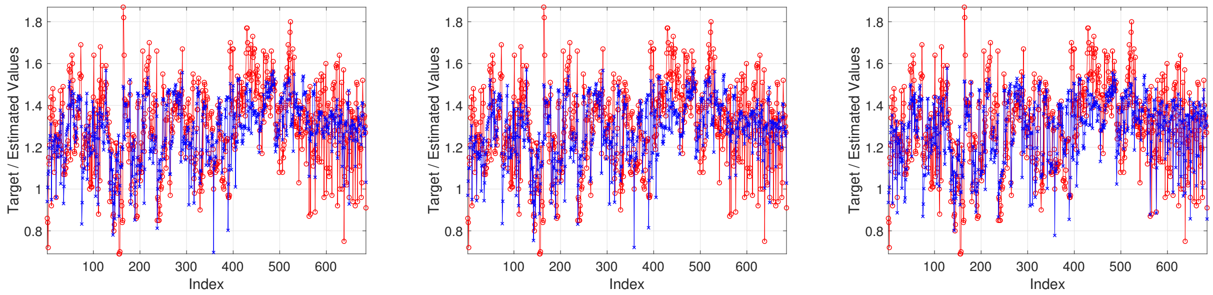
The next set of figures present a visual depiction of the behavior of the methods developed and is intended to complement the information provided in Table II.

The first set of images, constituted by Figure 5a, Figure 5b and Figure 5c, presents, respectively, the comparisons between target d and estimated values y for the LS, LS′, and the committee methods. This first set illustrates that there is indeed some overlap between the targets and the estimates, as should be expected. Figure 6 illustrates the difference between curves of the previous data. All plots are normalized for both axes in order to show that the committee errors are more restricted in range. Although we get a better understanding of the behavior of the methods it is still desirable to look at the data from another perspective.

The second set of pictures, encompassing Figure 7a, Figure 7b and Figure 7c, presents, respectively, the scatter plots between target d and estimated values y for the LS, LS′ and the committee methods. Notice that although there is some tendency for over- and under-estimation the committee method is the one that is able to provide the closest approximation.

We conclude by presenting a description of the histogram counts of the errors. The third set of images includes Figure 8a, Figure 9a, Figure 8b, Figure 9b, Figure 8c and Figure 9c, which present, respectively, histograms of both

[3] Note that the fixed number of epochs attenuates the overfitting issues due to the employed cross-validation method.

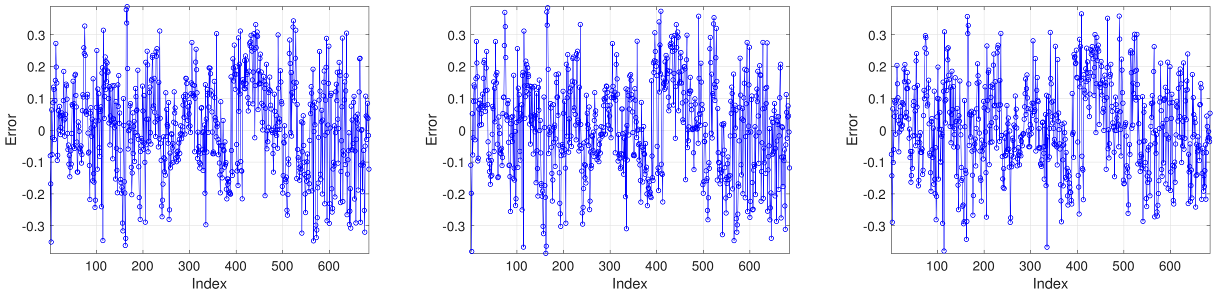


(a) Comparison against LS method.

(b) Comparison against LS′ method.

(c) Comparison against committee method.

Fig. 5. Comparison between target values d (red solid line) and the y values estimated (blue solid line) for each method.

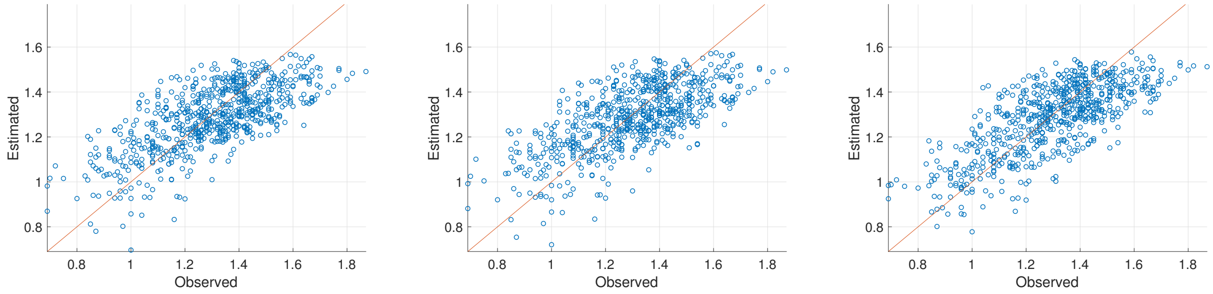


(a) LS method errors.

(b) LS′ method errors.

(c) Committee method errors.

Fig. 6. Errors for the different regressors.



(a) LS scatter plot.

(b) LS′ scatter plot.

(c) Committee scatter plot.

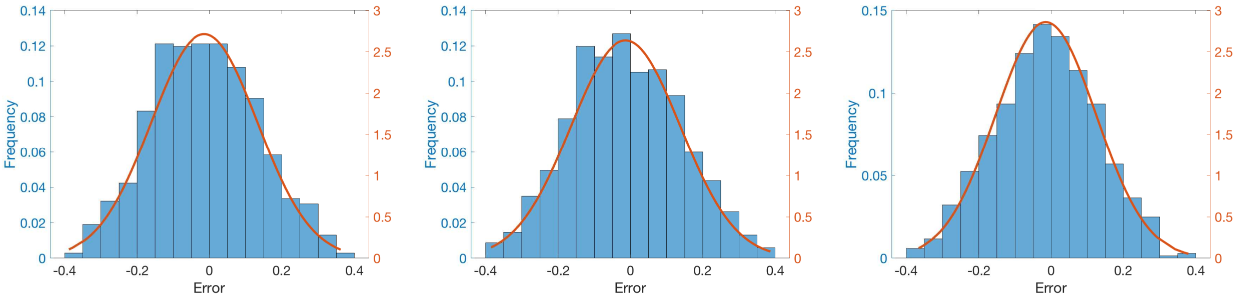
Fig. 7. Scatter plots for the different regression methods.

errors and absolute errors for the LS, LS′ and committee methods. Notice that non-absolute errors tend to follow a Gaussian distribution, with most errors accumulating around zero or its vicinity, which is the desired behavior. One-sample Kolmogorov-Smirnov test [53] reveals p-values of $0.7184, 0.6575$ and 0.9496 for, respectively, LS, LS′ and the committee. This reinforces the normality of the errors distribution of the committee.

VII. Final Remarks

In this work, we presented several techniques for estimating soil bulk density (D_s), a difficult metric to measure. We developed two methods based on least-squares regression, one using all the available correlation data and another discarding the two variables least correlated with D_{s}. An additional method was constructed based on a static committee of ANNs, which combines multiple regressors. Our results show that we are able to increase precision against the original least-squares method by using less information but also by employing a committee of regressors.

Our results emphasize that it is not clear the relationship between D_{s} and certain chemical attributes (such as pH, Sb and Al^{+3}). This result corroborates those in the existing literature, where a direct physical link between these properties is not clearly presented. In general, these attributes are inserted into the model due to their availability in most datasets, because of chemical attributes present low-cost acquisition [5].

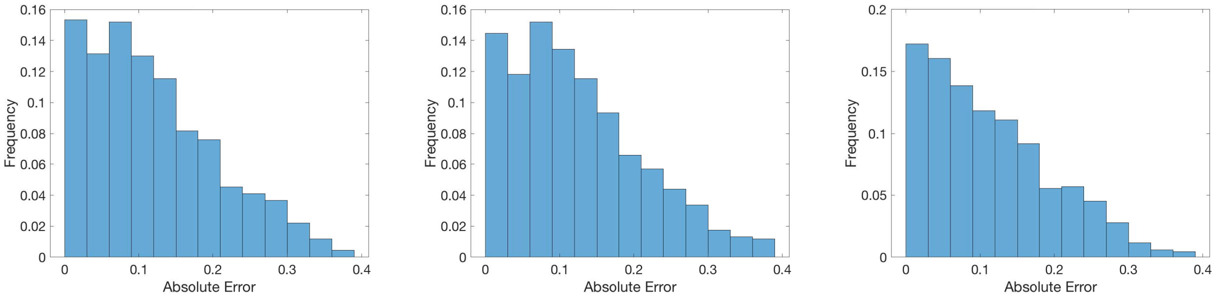


(a) LS errors($\mu = -0.0140, \sigma = 0.1470$).

(b) LS$'$ errors ($\mu = -0.0133, \sigma = 0.1512$).

(c) Committee errors ($\mu = -0.0144, \sigma = 0.1395$).

Fig. 8. Committee errors histogram (left y-axis) with superimposed Gaussian distribution (right y-axis).



(a) LS absolute errors.

(b) LS$'$ absolute errors.

(c) Committee absolute errors.

Fig. 9. Committee absolute errors histogram (left y-axis) with superimposed Gaussian distribution (right y-axis).

Gomes *et al.* [11] hypothesized that Al^{+3} and H availability increases as the pH is lowered, increasing soil porosity and decrease soil bulk density.

ACKNOWLEDGMENTS

This work was developed with the support of CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico (National Council for Scientific and Technological Development, in Brazil), CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Coordination for Enhancement of Higher Education Personnel, in Brazil), and FAPERJ - Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (Rio de Janeiro Research Support Foundation, in Brazil).

REFERENCES

[1] M. Budiman, A. B. Bratney, M. L. M. Santos, and H. G. dos Santos, *Reviso sobre funes de pedotransferłncia (PTFs) e novos mtodos de predio de classes e atributos do Solo*. Embrapa Solos, 2003.

[2] R. S. Boschi, F. F. Bocca, M. L. R. C. Lopes-Assad, and E. D. Assad, “How accurate are pedotransfer functions for bulk density for brazilian soils?,” *Scientia Agricola*, vol. 75, pp. 70–78, Jan 2018.

[3] V. M. Benites, P. L. O. A. Machado, E. C. C. Fidalgo, M. C. Coelho, and B. E. Madari, “Pedotransfer functions for estimating soil bulk density from existing soil survey reports in brazil,” *Geoderma*, vol. 139, pp. 90–97, 2007.

[4] M. I. Al-Qinna and S. M. Jaber, “Predicting soil bulk density using advanced pedotransfer functions in an arid environment,” *Transactions of the American Society of Agricultural and Biological Engineers*, vol. 56, pp. 963–976, Feb 2013.

[5] M. Bernoux, D. Arrouays, C. Cerri, B. Volkoff, and C. Jolivet, “Predicting soil bulk density using advanced pedotransfer functions in an arid environment,” *Soil Science Society of America Journal*, vol. 62, pp. 743–749, May 1998.

[6] T. Moreira, D. N. B. ao, D. B. Haddad, M. B. Ceddia, E. F. M. Pinheiro, and R. F. Oliveira, “A first approach using neural network to estimating soil bulk density of urucu basin in central amazon-brazil,” in *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 3236–3239, May 2017.

[7] L. Alletto and Y. Coquet, “Temporal and spatial variability of soil bulk density and near-saturated hydraulic conductivity under two constrated tillage management systems,” *Geoderma*, vol. 152, pp. 85–94, 2009.

[8] K. Sun, S. H. Huang, D. S. H. Wong, and S. S. Jang, “Design and application of a variable selection method for multilayer perceptron neural network with lasso,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, pp. 1386–1396, June 2017.

[9] W. J. Rawls, T. J. Gish, and D. L. Brakensiek, *Estimating Soil Water Retention from Soil Physical Properties and Characteristics*, vol. 16. Springer, New York, NY, 1991.

[10] R. W. J. Pachepsky, Y. A. and D. J. Timlin, *The Current Status of Pedotransfer Functions: Their Accuracy, Reliability, and Utility in Field and RegionalScale Modeling*, vol. 108. January 1999.

[11] A. Gomes, A. Ferreira, E. F. M. Pinheiro, M. Menezes, and M. Ceddia, “The use of Pedotransfer functions and the estimation of carbon stock in the Central Amazon region,” *Scientia Agricola*, vol. 74, pp. 450 – 460, 12 2017.

[12] A. A. Shalmani, M. S. Shahrestani, H. Asadi, and F. Bagheri, “Comparison of regression pedotransfer functions and artificial neural networks for soil aggregate stability simulation,” *World Applied Sciences Journal*, vol. 8, no. 9, pp. 1065–1072, 2010.

[13] L. R. Lado, M. Rial, T. Taboada, and A. M. Cortizas, “A pedotransfer function to map soil bulk density from limited data,” *Procedia Environmental Sciences*, vol. 27, pp. 45–48, 2015.

[14] S. J. Beutler, M. G. Pereira, T. W. S, M. D. Menezes, G. S. Valladares, and L. H. C. Anjos, “Bulk density prediction for histosols and soil

horizons with high organic matter content," *Revista Brasileira de Ciência do Solo*, vol. 41, p. 14, Jan 2017.

[15] S. Chen, A. C. R. Forgesa, N. P. A. Sabya, M. P. Martina, C. Walterb, and D. Arrouaysa, "Building a pedotransfer function for soil bulk density on regional dataset and testing its validity over a large area," *Geoderma*, vol. 312, pp. 52–63, Jan 2018.

[16] H. S. Barros and P. M. Fearnside, "Pedo-transfer functions for estimating soil bulk density in central amazonia," *Revista Brasileira de Ciência do Solo*, vol. 39, pp. 397–407, 2015.

[17] C. H. Sequeira, S. A. Wills, C. A. Seybold, and L. T. West, "Predicting soil bulk density for incomplete databases," *Geoderma*, vol. 213, pp. 64–73, Sept 2014.

[18] L. Wang, X. Y. Zhang, and C. Pan, "Msdlsr: Margin scalable discriminative least squares regression for multicategory classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, pp. 2711–2717, Dec 2016.

[19] L. Wang, S. Liu, and C. Pan, "Rodlsr: Robust discriminative least squares regression model for multi-category classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2407–2411, March 2017.

[20] G. S. Galloway, V. M. Catterson, C. Love, A. Robb, and T. Fay, "Modeling and interpretation of tidal turbine vibration through weighted least squares regression," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. PP, no. 99, pp. 1–8, 2017.

[21] X. Y. Zhang, L. Wang, S. Xiang, and C. L. Liu, "Retargeted least squares regression algorithm," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, pp. 2206–2213, Sept 2015.

[22] T. Strutz, *Data fitting and uncertainty: A practical introduction to weighted least squares and beyond*. Vieweg and Teubner, 2010.

[23] L. Wang and C. Pan, "Groupwise retargeted least-squares regression," *IEEE Transactions on Neural Networks and Learning Systems*, vol. PP, no. 99, pp. 1–7, 2017.

[24] D. B. Haddad, L. O. Nunes, W. A. Martins, L. W. P. Biscainho, and B. Lee, "Closed-form solutions for robust acoustic sensor localization," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 1–4, Oct 2013.

[25] R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye, *Probability and statistics for engineers and scientists*, vol. 5. Macmillan New York, 1993.

[26] C. Bras-Geraldes, A. Papoila, P. Xufre, and F. Diamantino, "Generalized additive neural networks for mortality prediction using automated and genetic algorithms," in *2013 IEEE 2nd International Conference on Serious Games and Applications for Health (SeGAH)*, pp. 1–8, May 2013.

[27] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li, "Multi-view gait recognition based on motion regression using multilayer perceptron," in *2010 20th International Conference on Pattern Recognition*, pp. 2186–2189, Aug 2010.

[28] D. B. L. Bong, J. Y. B. Tan, and K. C. Lai, "Application of multilayer perceptron with backpropagation algorithm and regression analysis for long-term forecast of electricity demand: A comparison," in *2008 International Conference on Electronic Design*, pp. 1–5, Dec 2008.

[29] C. A. C. Montaez, P. Fergus, A. Hussain, D. Al-Jumeily, B. Abdulaimma, J. Hind, and N. Radi, "Machine learning approaches for the prediction of obesity using publicly available genetic profiles," in *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 2743–2750, May 2017.

[30] N. A. Agana and A. Homaifar, "A deep learning based approach for long-term drought prediction," in *SoutheastCon 2017*, pp. 1–8, March 2017.

[31] A. Shalaginov, "Evolutionary optimization of on-line multilayer perceptron for similarity-based access control," in *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 823–830, May 2017.

[32] D. Kim, J. Kung, and S. Mukhopadhyay, "A power-aware digital multilayer perceptron accelerator with on-chip training based on approximate computing," *IEEE Transactions on Emerging Topics in Computing*, vol. 5, pp. 164–178, April 2017.

[33] M. Hashemi, "Reusability of the output of map-matching algorithms across space and time through machine learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, pp. 3017–3026, Nov 2017.

[34] S. S. Haykin, S. S. Haykin, S. S. Haykin, and S. S. Haykin, *Neural networks and learning machines*, vol. 3. Pearson Upper Saddle River, NJ, USA:, 2009.

[35] H. B. Demuth, M. H. Beale, O. De Jess, and M. T. Hagan, *Neural network design*. Martin Hagan, 2014.

[36] D. Yu and L. Deng, *Automatic speech recognition: A deep learning approach*. Springer, 2014.

[37] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural computation*, vol. 4, no. 1, pp. 1–58, 1992.

[38] Y. Liu, J. A. Starzyk, and Z. Zhu, "Optimized approximation algorithm in neural networks without overfitting," *IEEE Transactions on Neural Networks*, vol. 19, pp. 983–995, June 2008.

[39] Z. Lv, S. Luo, Y. Liu, and Y. Zheng, "Information geometry approach to the model selection of neural networks," in *First International Conference on Innovative Computing, Information and Control - Volume I (ICICIC'06)*, vol. 3, pp. 419–422, Aug 2006.

[40] C. Xiang, S. Q. Ding, and T. H. Lee, "Geometrical interpretation and architecture selection of mlp," *IEEE Transactions on Neural Networks*, vol. 16, pp. 84–96, Jan 2005.

[41] L. Holmstrom and P. Koistinen, "Using additive noise in backpropagation training," *IEEE Transactions on Neural Networks*, vol. 3, pp. 24–38, Jan 1992.

[42] S. Amari, N. Murata, K. R. Muller, M. Finke, and H. H. Yang, "Asymptotic statistical theory of overtraining and cross-validation," *IEEE Transactions on Neural Networks*, vol. 8, pp. 985–996, Sep 1997.

[43] G. N. Karystinos and D. A. Pados, "On overfitting, generalization, and randomly expanded training sets," *IEEE Transactions on Neural Networks*, vol. 11, pp. 1050–1057, Sep 2000.

[44] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*. CRC press, 1994.

[45] W. Yan, "Toward automatic time-series forecasting using neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, pp. 1028–1039, July 2012.

[46] J. Yang and S. Luo, "Hybrid committee machine for incremental learning," in *2005 International Conference on Neural Networks and Brain*, vol. 1, pp. 391–395, Oct 2005.

[47] W. C. Chung, C. Y. Chang, and C. C. Ko, "A svm-based committee machine for prediction of hong kong horse racing," in *2017 10th International Conference on Ubi-media Computing and Workshops (Ubi-Media)*, pp. 1–4, Aug 2017.

[48] M. Ferrer, M. de Diego, A. Gonzalez, and G. Piero, "Convex combination of affine projection algorithms," in *2009 17th European Signal Processing Conference*, pp. 431–435, Aug 2009.

[49] J. Arenas-Garcia, A. R. Figueiras-Vidal, and A. H. Sayed, "Mean-square performance of a convex combination of two adaptive filters," *IEEE Transactions on Signal Processing*, vol. 54, pp. 1078–1090, March 2006.

[50] P. J. Huber, "Robust statistics," in *International Encyclopedia of Statistical Science*, pp. 1248–1251, Springer, 2011.

[51] C. Bishop, "Pattern recognition and machine learning (information science and statistics), 1st edn. 2006. corr. 2nd printing edn," *Springer, New York*, 2007.

[52] G. Dreyfus, *Neural networks: methodology and applications*. Springer Science & Business Media, 2005.

[53] F. J. Massey Jr, "The kolmogorov-smirnov test for goodness of fit," *Journal of the American statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.