**Soil spectral library of Piauí State using machine learning for laboratory analysis in Northeastern Brazil**

Wanderson de Sousa Mendes(1)*, Cácio Luiz Boechat(2), Adriano Venicius Santana Gualberto(2), Ronny Sobreira Barbosa(2), Yuri Jacques Agra Bezerra da Silva(2), Paloma Cunha Saraiva(2), Antonny Francisco Sampaio de Sena(3) and Lizandra de Sousa Luz Duarte(3)

(1) OpenGeoHub Foundation, Wageningen, the Netherlands.  
(2) Universidade Federal do Piauí, Campus Professora Cinobelina Elvas, Bom Jesus, Piauí, Brasil.  

**ABSTRACT:** Soil chemical and physical analyses are the major sources of data for agriculture. However, traditional soil analyses are time-consuming, not cost-efficient, and not environmentally friendly. An alternative to traditional soil analyses is soil spectroscopy. This technique is a low-cost and quick analytical method, which can be implemented in a laboratory and/or in-situ. Nevertheless, some spectrometers are expensive and do not contemplate the entire spectrum. Despite this limitation, the main objective of the study was to create a soil spectral library of the Piauí State using only the 1000–2500 nm range. In this sense, it was evaluated and standardized the soil spectral library by accessing the combination of smoothing, standard normal variate, continuum removal, and Savitzky-Golay derivative spectral preprocessing procedures with partial least squares, random forest, and cubist machine learning algorithms. It was collected 262 geo-referenced soil samples at the layer of 0.00–0.20 m across the entire Piauí State, representing most of its soil variability. The soil properties evaluated were pH(H₂O), sand, clay, and soil organic carbon (SOC) contents. This study demonstrated that the Standard Normal Variate was one of the most promising preprocessing procedures to improve model predictions for pH(H₂O), sand, and clay. For SOC and pH, the best overall results were without preprocessing the soil spectra. Moreover, the cubist model was the most accurate in predicting soil properties. Finally, our study showed evidence of the potential and feasibility of using this soil spectral library to estimate soil properties such as pH(H₂O), sand, clay, and SOC.

**Keywords:** soil spectroscopy, soil predictions, shortwave infrared, NIR.
INTRODUCTION

Soil chemical and physical analyses are the principal sources of data for agriculture. This information allows to better manage the carbon cycle (Guevara et al., 2020), plants growth (Soong et al., 2020), soil acidity and fertility (Cargnelutti Filho et al., 1996), and is part of the soil security framework (Amundson et al., 2015; Bennett et al., 2019). However, those analyses are performed in traditional soil laboratories that produce potential toxic residuals for the environment. For instance, one unique soil sample produces ~1.4 g of Cr₂O₇²⁻ and Fe(NH₄)₂(SO₄)₂·6H₂O, and ~5 mL of H₂SO₄ for the analysis of soil organic carbon (SOC) only. In 2019, the Official Network of Soil and Plant Tissue Analysis Laboratories in the Brazilian states of Rio Grande do Sul and Santa Catarina performed 286,659 traditional soil analyses (ROLAS, 2019). A rough calculation shows that those analyses for SOC produced approximately 401.3 kg and 1,433.3 L of toxic residuals per year, which were dumped in the sewerage network (Afonso et al., 2003). Other issues of traditional soil analyses are time-consuming and not cost-efficient (Fang et al., 2018; Jaconi et al., 2019; Borges et al., 2020).

An alternative to traditional soil analyses is soil spectroscopy, which consists of using the emission and absorption of light and other radiation (electromagnetic radiation) by soils (matter). This technique is a low-cost and quick analytical method, which can be implemented in a laboratory and/or in-situ. Nevertheless, some spectrometers are expensive and do not contemplate the entire spectrum. Despite this limitation, it is important to evaluate the capacity of using specific ranges of the spectrum to quantify soil properties. In the last few decades, the diffuse reflectance has arisen and matured as a notable identifier of physical, chemical, and mineralogical components of soils (Stoner and Baumgardner, 1981; Ben-Dor and Banin, 1995; Nocita et al., 2015; Rossel et al., 2016; Demattê et al., 2019). The soil spectroscopy mainly encompasses the visible (Vis, 350–700 nm) and near-infrared (NIR, 700–2500 nm) ranges of the electromagnetic spectrum. In remote sensing, the NIR is usually split into two ranges because of the satellite sensors (Ogen et al., 2017; Crucil et al., 2019; Mendes et al., 2019) as NIR (700–1000 nm) and shortwave infrared (SWIR, 1000–2500 nm) (López-Maestresalas et al., 2016; Wilczyński et al., 2016). The composition presented in this study fixes in the SWIR region because our sensor only captures this range of the electromagnetic spectrum.

Even though there is the Brazilian Soil Spectral Library (Demattê et al., 2019), which was a national effort to build predictive models at distinct geographic levels, acquiring precise spectroscopic models continues a demanding assignment (Grunwald et al., 2015; Moura-Bueno et al., 2020), mainly at downscales. For instance, the number of samples from Piauí State in that soil spectral library was 67 units, which is a low sampling density to calibrate and validate a state quantitative model such as the area of Piauí State (~251,000 km²). Those authors pointed out that some predictive models presented worse results because of the small number of samples at the state level. Thus, there is a gap to be fulfilled at regional and local scales developing specific soil spectral libraries that better describe quantitatively and qualitatively the physical, chemical, and mineralogical soil properties. Therefore, the main objective was to develop a specific spectral library for Piauí State and; in this sense, this study aimed to (i) evaluate the potentiality of using only the SWIR range to build predictive models and (ii) access which is the best combination of the spectral preprocessing procedure (smoothing, standard normal variate, continuum removal, and Savitzky-Golay derivatives) with three machine learning regression methods (partial least squares, random forest, and cubist) for pH(H₂O), sand, clay, and SOC at the regional scale.

MATERIALS AND METHODS

Study area, soil sampling, and analyses

The study area covers approximately 251,000 km² and is the entire extent of Piauí State in the Northeastern Brazil. The two ecotones of the region are Caatinga and Cerrado biomes.
(Figure 1). The first one is characterized by large depressions, plateaus, mainly xerophytic flora, and is not as ecologically diverse as other Brazilian ecotones (Oliveira et al., 2019). The Cerrado biome is typified by highlands, savannah flora, and biodiversity. In both biomes, the main soil types, according to the Brazilian Soil Classification System (Santos et al., 2018), are Latossolos, Argissolos, Cambissolos, Neossolos Regolíticos, and Neossolos Quartzarênicos, which correspond to Ferralsols, Acrisols, Cambisols, Regosols, and Arenosols (IUSS Working Group WRB, 2015), respectively. Approximately 85 % of the Piauí's territory are constituted by phanerzoic sedimentary rocks, and the remaining 15 % are occupied by metamorphic and igneous rocks. Consequently, the soil types are predominately Latossolos (Ferralsols), Neossolos (Regosols and Arenosols), and Argissolos (Acrisols). The region is classified as tropical wet with dry winter (Aw) in the Cerrado biome, as hot semi-arid with dry winter (BShw) in the Caatinga biome, and as hot and humid tropical with dry winter on the Koppen classification system displaying the average annual temperature of 26.7 °C (Alvares et al., 2013).

We collected 262 geo-referenced soil samples at 0.00–0.20 m layer based on geology, relief, and under native vegetation or non-cultivated areas, which comprise areas

![Figure 1. Methodological sequence of the research.](image-url)
of low impact or landscapes with low or non-anthropic interference. Those samples covered most of the variability of the soil parent materials and types. To ensure a good representation, each soil sample was composed of five sub-samples collected 4 m from the central geo-referenced point using an auger, sieve, and bucket of stainless steel. The soil chemical and physical analyses were carried out in soil samples on oven-dried (48 h at 50 °C), ground, and sieved (<2.0 mm mesh). The soil pH was measured in an extract at 1:2.5 soil/water ratio, which was stirred and kept at rest for 1 hour at room temperature. The soil particle size, clay and sand, was determined by the Bouyoucos densimeter method (Beretta et al., 2014; Day, 2015). The soil organic carbon was determined by the Walkley-Black method by oxidation of the carbon in 10 mL of K₂Cr₂O₇ 0.0667 mol L⁻¹ via wet and subsequent titration with (NH₄)₂(SO₄)₂ 0.102 mol L⁻¹ (Walkley and Black, 1934). Following the recent recommendation stated by Minasny et al. (2020), drawing briefly the history of the van Bemmelen’s factor (e.g., 1.724) to convert organic carbon into organic matter, we modeled the soil organic carbon instead of soil organic matter. Those authors pointed out that it would be further meticulous in reporting soil organic carbon.

Soil spectral acquisition and preprocessing

The soil spectral data were acquired using a multi-range Fourier Transform-Infrared/ Near-Infrared spectrometer (FT-IR/NIR, PerkinElmer Inc., USA) in a range of 1000 to 2500 nm with a spectral resolution of 0.5 nm. The ground and sieved (<2.0 mm mesh) soil samples were placed in a petri dishes for homogeneous arrangement allowing a flat surface for performing the spectral reading. The detector type of the sensor is DTGS, which the signal is generated by the change in temperature altered by absorption of the infrared radiation operated at room temperature (Figure 1). The sensor covers an area of approximately 2 cm³ in the center of each sample. A white plate called Spectralon, which represents a ~100 % reflectance pattern, was used to calibrate the sensor. This procedure was performed at the beginning of each reading to standardize and equalize the spectral measurements.

Beyond the raw spectra, four spectral preprocessing approaches were applied to mathematically correct measurements and noisy. Moreover, these four techniques were selected because of their ability to provide tuning models with high accuracy. The first one was the smoothing, which is a straightforward moving average of spectral data using a convolution function. We used an 11-nm search window to smooth our soil spectra data (Figure 2a). The second preprocessing was the Standard Normal Variate (SNV), which removes scatter in the spectral data. It normalizes each spectrum considering its mean value and the standard deviation (Figure 2b) (Barnes et al., 1989). The third was continuum removal (CR), which eliminates the continuum aspects of spectra and frequently isolates specific absorption characteristics (Figure 2c). The CR finds points placing on the convex hull of 100 % of reflection, joins the points by linear interpolation and standardizes the spectrum by splitting the input data by the interpolated line (Clark and Roush, 1984). Last but not least, the Savitzky-Golay (SG) smoothing minimizes the set of variation and enhances resolutions of spectral peak characteristics (Savitzky and Golay, 1964). We used SG with the first derivative with a first-order polynomial and window size of 5-nm (Figure 2d). The soil spectra preprocessing was performed using the Alradspectra graphical user interface in R programming (Dotto et al., 2019). The setting parameters for SNV and SG were selected after we performed the initial tests.

Spectral modeling and model assessment

The spectral modeling was performed employing the partial least squares regression (PLSR), random forest, and cubist machine learning methods in the R programming (R Development Core Team, 2020). The dataset was randomly split into calibration and validation. The models were calibrated using 80 % of the dataset and implemented in R
package “caret” (Kuhn, 2008). The tuning parameters for each algorithm are described in Table 1. The optimization was performed using a 10-fold repeated cross-validation method, executed ten times for each model. The best models were selected based on the lowest RMSE and highest $R^2$ of calibration and validation dataset.

The PLSR joins features from multiple regression, and principal components analysis diminishing complex spectral matrix solving single- and multi-label learning issues and is widely applied in soil spectroscopic modeling (Ergon, 2013). The adjustable parameter of PLSR is the selection of principal components that explain 95% of the variance of the latent variables (number of principal components). The cubist model is based on the construction of regression trees, wherein the set of response variables are split into subsets by similarity. The partitions are defined by if-then conditions that generate a sequence of rules. The regulating arguments of the cubist model are interactive model trees created in sequence (committee) and the nearest neighbours, which determine the average of a training set points. Detailed information is described in Quinlan (1993). Last but not least, the random forest computes a mean decrease impurity estimating the feature importance. The adjustable parameters are the number of trees and the number of variables randomly sampled as candidates at each split (Breiman, 2001).

The model evaluation was done using 20% of the observations left for external validation and the performance of the prediction models assessed employing the root mean square error (RMSE; Equation 1), coefficient of determination ($R^2$; Equation 2), the ratio of performance to deviation (RPD; Equation 3), and the ratio of performance to interquartile distance (RPIQ; Equation 4).
\[ \text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}} \]  
\[ R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \]  
\[ \text{RPD} = \frac{SD}{\text{RMSE}} \]  
\[ \text{RPIQ} = \frac{IQ}{\text{RMSE}} \]

In which \( \hat{y} \) is the predicted value; \( \bar{y} \) is the mean observed value; \( y \) is the observed values; \( n \) is the number of samples with \( i = 1, 2, 3, ..., n \); \( SD \) is the standard deviation of the observed values; and \( IQ \) is the interquartile range of data distribution calculated by the difference between the 3rd and 1st quartile. The RPD values were used just for comparison to the literature. Bellon-Maurel et al. (2010) pointed out that this index's thresholds are arbitrary and without statistical or utilitarian basis.

**RESULTS**

**Soil descriptive and spectral qualitative analyses**

Table 2 presents the distribution of pH(H2O), sand, clay, and soil organic carbon in the study area. The pH(H2O) values varied from extremely acid (3.5–4.5) to strongly alkaline...
However, the more predominant pH(H2O) values are around 4.52. This pattern is an intrinsic characteristic of the region. The clay content presented values ranging from 107.00 to 508.00 g kg\(^{-1}\), with a median of 163.40 g kg\(^{-1}\). This shows the predominance of sandy soils. The soil organic carbon (SOC) content was up to 2.5 dag kg\(^{-1}\). The skewness and kurtosis indices for all properties demonstrated a satisfactory normal distributed dataset. These explained patterns help to understand the importance of regional/local datasets to build a trustable and representative spectral library.

The soil spectra have both information qualitative and quantitative. The qualitative aspects of the soil spectra are on the absorption and reflectance intensities called valleys and peaks, respectively. Figure 3 shows the median spectral curves of the clay, pH(H2O), and SOC groups. The clay contents were grouped in clay (350–600 g kg\(^{-1}\)), loam sand (150–350 g kg\(^{-1}\)), sand (100–150 g kg\(^{-1}\)), and very sand (<100 g kg\(^{-1}\)). Qualitatively, there is a slight difference between sandy and very sandy soil spectral signature, which could be grouped into one class in our specific case. Clayey soils absorb more electromagnetic magnetic energy than sandy soils (Figure 3a). The SOC were divided into three classes: low (<1 dag kg\(^{-1}\)), medium (1–2 dag kg\(^{-1}\)), and high (>2.5 dag kg\(^{-1}\)). We suggested these classes to better distinguish SOC contents in our dataset because our spectrometer has only the SWIR window (Figure 3b). The arrangement for the pH(H2O) values was: very low (3–4.5), low (4.5–5.5), good (5.5–6), high (6–7), and very high (>7). Grouping the pH(H2O) values into these classes, we could qualitatively discriminate the pH(H2O) content at 1900 nm (Figure 3c).

**Soil spectral quantification**

Analyzing the Pearson’s correlation index (Figure 3d), the clay and SOC contents had inverse correlation up to -0.5. The pH(H2O) and sand values had a positive correlation up to 0.5 with the reflectance. The predictive metrics of the models was accessed by checking the external validation dataset left out, which simulate in-situ new sampling (Table 3). Assessing the models for pH(H2O) prediction, the SNV displayed as the best preprocessing procedure among other models (Table 3). The cubist model (RMSE = 0.54; \(R^2 = 0.72\); and RPIQ = 1.62) was the best fitted comparing to PLSR and RF. The spectral modeling of pH(H2O) that showed higher RMSE (1.02), low \(R^2\) (0.26), and RPIQ (0.92) was the RF using the smoothing preprocessing technique.

**Table 2.** Descriptive statistics of pH in water, sand, clay, and organic carbon contents in the Piauí State, Brazil

<table>
<thead>
<tr>
<th>Parameters</th>
<th>pH(H2O)</th>
<th>Sand</th>
<th>Clay</th>
<th>Soil Organic Carbon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td>3.11</td>
<td>188.90</td>
<td>1.00</td>
<td>0.01</td>
</tr>
<tr>
<td>1st Quartile</td>
<td>3.79</td>
<td>645.30</td>
<td>107.00</td>
<td>0.37</td>
</tr>
<tr>
<td>Median</td>
<td>4.52</td>
<td>761.40</td>
<td>163.40</td>
<td>0.64</td>
</tr>
<tr>
<td>Mean</td>
<td>4.68</td>
<td>729.00</td>
<td>171.40</td>
<td>0.88</td>
</tr>
<tr>
<td>3rd Quartile</td>
<td>5.33</td>
<td>844.10</td>
<td>220.40</td>
<td>1.36</td>
</tr>
<tr>
<td>Maximum</td>
<td>8.57</td>
<td>990.10</td>
<td>508.00</td>
<td>2.50</td>
</tr>
<tr>
<td>CV (%)</td>
<td>22.78</td>
<td>20.87</td>
<td>54.88</td>
<td>78.82</td>
</tr>
<tr>
<td>SD (±)</td>
<td>1.06</td>
<td>152.10</td>
<td>94.00</td>
<td>0.69</td>
</tr>
<tr>
<td>Skewness</td>
<td>0.94</td>
<td>-0.67</td>
<td>0.25</td>
<td>0.79</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>0.57</td>
<td>-0.29</td>
<td>-0.43</td>
<td>-0.42</td>
</tr>
</tbody>
</table>

SD: Standard deviation; CV: coefficient of variation; The soil pH was measured in an extract at 1:2.5 soil/water ratio, which was stirred and kept at rest for one hour at room temperature. The soil particle size, clay, and sand, was determined by the Bouyoucos densimeter method (Beretta et al., 2014; Day, 2015). The soil organic carbon was determined by the Walkley-Black method by oxidation of the carbon in 10 mL of K_2Cr_2O_7 0.0667 mol L\(^{-1}\) via wet and subsequent titration with (NH_4)_2(SO_4)\_2.6H_2O 0.102 mol L\(^{-1}\) (Walkley and Black, 1934).
Overall, the SNV data was better fitted in all models for sand predictions. Using the PLSR model, the RMSE, $R^2$, RPD, and RPIQ values were 128.10 g kg$^{-1}$, 0.19, 1.13, and 0.91, respectively (Table 3). The results using RF were 110.10 g kg$^{-1}$, 0.43, 1.31, and 1.06 for RMSE, $R^2$, RPD, and RPIQ. However, the best fitted model for sand predictions was using the cubist algorithm (RMSE = 98.7 g kg$^{-1}$, $R^2$ = 0.52, RPD = 1.46, and RPIQ = 1.18). Applying the PLSR, the SGD dataset better fitted the clay predictions than other datasets preprocessed or not. For the other two algorithms, the SNV dataset was still the most advised dataset to be used in the calibration of RF and cubist models. However, the

Figure 3. Median of spectral curves grouped by clay (a), soil organic carbon (b), and pH(H$_2$O) (c) contents. Pearson correlation index between the soil properties and spectral reflectance in the shortwave infrared (d).
best models for predicting clay content were PLSR using SGD dataset (RMSE = 77.30 g kg\(^{-1}\); \(R^2 = 0.21\); and RPIQ = 0.76), and cubist using SNV dataset ((RMSE = 80.20 g kg\(^{-1}\); \(R^2 = 0.25\); and RPIQ = 0.73). The SGD dataset fitted more accurate SOC predictions than using the raw and other preprocessed datasets through PLSR models. However, the most overall accurate result was that using raw dataset with cubist models (RMSE = 0.49 dag kg\(^{-1}\); \(R^2 = 0.42\); and RPIQ = 0.41).

**DISCUSSION**

The more predominant pH(H\(_2\)O) values in the study area were around 4.52, which is considered a very strongly acid environment, following the classification proposed by van Raij et al. (2001). Leite et al. (2010) analyzed the chemical properties of different soil-uses in Piauí State’s savannah, and presented a dataset with pH(H\(_2\)O) values between 4.17 and 5.04 for native Cerrado forest at 0.00–0.40 m layer. This pattern was also similar in our dataset, including the Caatinga Biome. The predominance of sandy soils influences the soil organic carbon pools, reflecting in the water retention and other soil properties (Syers et al., 1970; Rawls et al., 2003; Vinther et al., 2006). It is not surprising that soil organic carbon (SOC) values are usually low. The symmetric parameters (skewness and kurtosis) for the properties demonstrated a satisfactory normal distributed dataset, which is not normally easy to find in a natural environment (Vereecken et al., 2016).
Vis-NIR field measurements can be influenced by water at 1400 and 1900 nm (Rossel et al., 2017). However, after the samples are oven-dried and Vis-NIR laboratory readings conducted, these two absorption features reveal the presence of phyllosilicates such as 2:1, 1:1 or interstratified minerals (Rossel et al., 2016; Demattê et al., 2017; Di Iorio et al., 2019). Evaluating our datasets’ spectral features qualitatively, they corroborated the fact that the soils in Piauí are rich in kaolinite and gibbsite (Figure 3a). These spectral features are presented by the characteristic doublet absorption valleys at 1400 and 2200 nm. Marques et al. (2019) evaluated the qualitative and quantitative spectral features of 16 soil profiles and pointed out that spectral features and a flat shape at 2200 and 2265 nm indicate the presence of kaolinite and gibbsite. This is caused due to the O-H bonds at 1400 nm and combined O-H stretch with Al-OH lattice at 1900 nm vibrations of water in the structure of phyllosilicates or absorbed on mineral surfaces (Clark et al., 1990). Kaolinitic and gibbsitic soils are the most dominant in tropical environments (Schwertmann and Herbillon, 2015), which corroborates our findings. Demattê et al. (2014) built a qualitative soil spectral library, and highlighted that the morphological features to identify organic matter are between 350 and 1100 nm, which comprises the visible up to the shortwave infrared ranges. This same trend was pointed out by Rossel et al. (2016), who created the global soil spectral library. The pH(H₂O) does not have active spectral features, but it is still measurable in the Vis-NIR because of the intrinsic relationship with the primary constituents of soils such as clay (Ge et al., 2020). As explained before, the absorption features at this wavelength present a H-O-H vibration of water that was in the mineral structure or absorbed on its surfaces (Rossel et al., 2016).

The inverse correlation between the SWIR spectral reflectance values and clay and SOC (Figure 3d) corroborated with the synthesis of high clay and SOC contents display high absorption spectral behavior (Clark et al., 1990; Demattê et al., 2019; Liu et al., 2020). The positive correlation of pH(H₂O) with the SWIR spectrum was because of the water molecular vibration, and the sand was caused due to the higher presence of quartz. Soil samples with high sand content present high reflectance (Demattê et al., 2018, 2019). Rossel et al. (2017), monitoring pH and SOC stocks in Northern New South Wales-Australia, obtained RMSE of 0.70 and 0.41 % and R² values of 0.71 and 0.81, respectively, using the Cubist machine learning algorithm with 317 samples to creating a local library for a 600 hectare. Demattê et al. (2019) produced the Brazilian Soil Spectral Library and evaluated some preprocessing and quantitative models for soil spectra at State, Region and National scales, found RMSE of 1.38 and R² of 0.29, and RPIQ of 1.68 for pH predictions using the cubist model and the CR preprocessing procedure in the Piauí State. Thus, our regional soil spectral library proves to be more accurate because of the sampling density and distribution influence by the greater representation of soils in the landscape. For the prediction models of sand, even though the R² and RPIQ values were lower than those presented by Demattê et al. (2019) (RMSE = 169.40 g kg⁻¹; R² = 0.80; and RPIQ = 3.83), the RMSE was lower in our study. The lower results in our findings for those parameters (e.g., R² and RPIQ) could be explained by the fact the equipment used in this study does not have the Vis-NIR window from 350–1000 nm. This spectral range strongly helps to identify the oxides, soil color and particles, and SOC (Clark et al., 1990; Ben-Dor and Banin, 1995; Ben-Dor et al., 2015; Nocita et al., 2015).

As it was for sand predictions, the prediction models of clay presented lower RMSE comparing to those findings reported by Demattê et al. (2019), which presented RMSE, R², and RPIQ values of 105.00 g kg⁻¹, 0.5, and 1.58, respectively. Our study presented better predictive metrics (e.g., RMSE) because the reduction in the number of samples undesirably interferes with the model performance for areas with high spectral and pedological variation (Moura-Bueno et al., 2020). Those authors had only 67 observations from Piauí State and thus, presented higher R² and RMSE. To improve the predictive power, it is probably necessary to add the Vis-NIR (350–1000 nm) or increase sampling density. In our specific case, better metric parameters of the clay and sand predictions
could be achieved by increasing sampling density. Most of the studies reported the soil organic matter or carbon predictions using Vis-NIR spectroscopy with transformed data by van Bemmelen’s factor (Rossel et al., 2016; Demattê et al., 2019). Rossel et al. (2006) reported RMSE values of 0.36 dag kg\(^{-1}\) for organic carbon predictions in the Vis-NIR range (350–1050 nm). As mentioned before, the most suitable ranges to predict SOC is in the Vis-NIR (350–1050 nm) and potentially in the MIR range (2500–25000 nm), as reported by Salazar et al. (2020) and Silvero et al. (2020). However, we found RMSE of 0.49 dag kg\(^{-1}\) and R\(^2\) of 0.42 using only the SWIR window and RMSE of 0.64 dag kg\(^{-1}\) using the same preprocessing and modeling procedure that those authors (e.g., Cubist and CR). Thus, in our study, we proved that if the sensor has a limited range of the Vis-NIR spectrum, it is still viable to use only the SWIR range to quantify some soil attributes. The raw data preprocessed via SNV improved spectral models of pH, sand, and clay predictions. Nevertheless, the best overall results were without preprocessing the soil spectra for SOC and pH. The most accurate model was the cubist, and that is the most applied model in the literature for soil predictions using Vis-NIR spectroscopy (Rossel et al., 2016; Demattê et al., 2019). As proved by Ng et al. (2019), soil spectral libraries with up to 2000 observations fit better the models using cubist and over 2000 observations it suggests to use a convolutional neural network. The performance overall results for pH(H\(_2\)O), sand, clay, and SOC were in according to those ranges reported by Soriano-Disla et al. (2014), Rossel et al. (2016), and Demattê et al. (2019).

**CONCLUSION**

This study demonstrated that the Standard Normal Variate was one of the most promising preprocessing procedures to improve model predictions for pH(H\(_2\)O) and sand, and SGD for clay. The best overall results were without preprocessing the soil spectra for SOC and pH. Moreover, the cubist model was the most accurate in predicting soil properties. Finally, our study showed evidence of the potential and feasibility of using only the SWIR spectral window to estimate soil properties such as pH(H\(_2\)O), sand, clay, and SOC. The regional soil spectral preprocessing recommendations for SWIR spectroscopy laboratory analysis to be applied in northeastern Brazil is to preprocess the raw/original data using SNV for pH, sand, and clay, and use the raw/original data for SOC. Furthermore, the main machine learning algorithm suggested is cubist. However, as the spectral library increases the number of observations, it is advised to perform a new evaluation using the convolutional neural network.

**ACKNOWLEDGMENTS**

This study was supported by the Brazilian National Research and Development Council - CNPq (Process Number: 409398/2016-0) and PQ Fellowship (CNPq Process Number: 303952/2017-2).

**AUTHOR CONTRIBUTIONS**

**Conceptualization:** Wanderson de Sousa Mendes (lead) and Cácio Luiz Boechat (supporting).

**Methodology:** Wanderson de Sousa Mendes (lead) and Cácio Luiz Boechat (supporting).

**Software:** Wanderson de Sousa Mendes (lead).

**Validation:** Wanderson de Sousa Mendes (lead).

**Formal analysis:** Wanderson de Sousa Mendes (lead).
Investigation: Antonny Francisco Sampaio de Sena (equal) and Lizandra de Sousa Luz Duarte (equal).

Resources: Cácio Luiz Boechat (lead).

Data curation: Wanderson de Sousa Mendes (lead).

Writing - original draft: Wanderson de Sousa Mendes (lead), Adriano Venicius Santana Gualberto (equal), and Cácio Luiz Boechat (equal).

Writing - review and editing: Wanderson de Sousa Mendes (equal), Cácio Luiz Boechat (equal), Adriano Venicius Santana Gualberto (equal), Ronny Sobreira Barbosa (equal), Yuri Jacques Agra Bezerra da Silva (equal), Paloma Cunha Saraiva (equal), Antonny Francisco Sampaio de Sena (equal), and Lizandra de Sousa Luz Duarte (equal).

Visualization: Wanderson de Sousa Mendes (equal), and Cácio Luiz Boechat (equal).

Supervision: Wanderson de Sousa Mendes (lead).

Project administration: Cácio Luiz Boechat (lead).

Funding acquisition: Cácio Luiz Boechat (lead).

REFERENCES


