Geoderma 337 (2019) 565–581

Contents lists available at ScienceDirect

Geoderma

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

Stratification of a local VIS-NIR-SWIR spectral library by homogeneity criteria yields more accurate soil organic carbon predictions

Jean Michel Moura-Bueno[a,*], Ricardo Simão Diniz Dalmolin[a], Alexandre ten Caten[b], André Carnieletto Dotto[c], José A.M. Demattê[c]

[a] Department of Soil, University Federal of Santa Maria, Av. Roraima, 1000, Building 42, Room 3314, 97105-900 Santa Maria, RS, Brazil

[b] Department of Agriculture, Biodiversity and Forest, Federal University of Santa Catarina, Rod. Ulysses Gaboardi, km 3, P. O. Box 101, 89520-000, Curitibanos, SC, Brazil

[c] Department of Soil Science, College of Agriculture Luiz de Queiroz, University of São Paulo, Av. Pádua Dias 11, Portal Box 9, 13418-900 Piracicaba, SP, Brazil

ARTICLE INFO

Handling Editor: Alex McBratney

Keywords:
Hyperspectral remote sensing
Soil carbon
Pre-processing transformations
Multivariate calibration
Stratification of spectral library

ABSTRACT

Considering the hypothesis that the predictive capacity of models is tied to soil characteristics, the stratification of a spectral library into groups is a strategy to improve the accuracy of the predictions. Thus, the objective of this study was to i) characterize and identify differences among spectra obtained for subtropical soils samples, ii) evaluate different pre-processing techniques and multivariate methods to propose SOC prediction models from the spectral data and iii) evaluate the performance of SOC prediction models calibrated from the stratification of a local library. A local spectral library of soils (n = 841 samples) was used in the *Planalto* region of the State of Rio Grande do Sul, Brazil. Soil classes that occur in the area are: Rhodic Ferralsol (FR) and Dystric Gleysol (GL). Land uses are: native forest (NFo), native field (NFi) and crops in no-tillage system (CTS). SOC was determined via wet combustion with sulphochromic solution. Spectral reflectance measurements were performed in the laboratory with a spectroradiometer in the range of 350–2500 nm. Six pre-processing techniques were applied to the spectra (including derivatives, normalization and non-linear transformations) and four multivariate calibration methods, namely, partial least squares regression (PLSR), multiple linear regression (MLR), support vector machines (SVM) and random forest (RF), were used with the objective of identifying the best combination to predict SOC. After determining the best combination, the spectral library was stratified into groups based on soil class, land use, sample layer and spectral characteristics. The models were built with 70% of the samples for calibration and 30% for independent validation. The coefficient of determination (R^2_v), root mean square error ($RMSE_v$) and ratio of performance to interquartile range ($RPIQ_v$) of the independent validation were used to evaluate the performance of the models. The spectral curves presented different absorption characteristics in relation to soil classes and land uses. SGD pre-processing technique had the highest R^2_v and $RMSE_v$ values for all models. Among the multivariate methods, PLSR had the best performance for SOC prediction for the total set of samples (R^2_v = 0.74, $RMSE_v$ = 0.52% and $RPIQ_v$ = 2.23), followed by models SVM, MLR, and RF. The FR-CTS (n = 445) group showed the best model performance after stratification, with R^2_v = 0.82, $RMSE_v$ = 0.29% and $RPIQ_v$ = 2.60. For some stratified groups, the use of a smaller number of samples to build the model reduced the performance of the models, suggesting that one must be careful when working with small datasets. This study highlights the potential for the application of VIS-NIR-SWIR spectroscopy as a reliable and economical tool to quantify SOC concentrations for subtropical soils with high levels of iron oxides and clay on a local scale. Predictive models can be improved when the variation in soil characteristics is considered, underscoring the need for a preliminary study examining the grouping of the sample set to validate the use of local spectral libraries for the prediction of soil properties.

1. Introduction

Intense modifications in land use, climate change and the demand to simulate future scenarios are reflected by an increased interest in soil spatial information (McBratney et al., 2014; Amundson et al., 2015). Thus, studies in soil science have focused on developing methods that

* Corresponding author.
E-mail address: bueno.jean1@gmail.com (J.M. Moura-Bueno).

https://doi.org/10.1016/j.geoderma.2018.10.015
Received 10 November 2017; Received in revised form 26 September 2018; Accepted 6 October 2018

enable predictions at local, regional, national, and global scales of a wide range of soil attributes (Grunwald et al., 2015). Soil organic carbon (SOC) is the attribute responsible for a series of functions in the environment, especially in the global carbon cycle (Stockmann et al., 2015). Consequently, it is necessary to generate reliable information about SOC. Nevertheless, mapping this attribute demands time and money because of the high sampling intensity required (Vasques et al., 2010). In addition, the diversity of environmental and anthropic factors also influences its distribution in the pedosphere. Therefore, faster and less expensive methods to quantify SOC are required.

The technique of reflectance spectroscopy in the visible (VIS – 350–700 nm), near-infrared (NIR – 700–1100 nm) and shortwave infrared (SWIR – 1100–2500 nm) ranges and the organization of soil spectral information in spectral libraries are potential tools to capture the pedologic variability (Dalmolin et al., 2005; Viscarra Rossel et al., 2016; Demattê et al., 2017) and quantify multiple soil attributes (Ben-Dor and Banin, 1995; Nocita et al., 2015; Guerrero et al., 2016; Dotto et al., 2017; Pinheiro et al., 2017; Vašát et al., 2017). Both approaches have a series of advantages when compared to traditional methods to determine SOC, such as the following: quick, cost-effective, and environmentally friendly. These attributes have been confirmed in a study by O'Rourke and Holden (2011), who evaluated the cost and time involved in SOC analysis between the Walkley-Black and VIS-NIR-SWIR spectroscopy techniques. The authors concluded that this technique is 10 times less expensive and associated with more rapid analyses, especially when evaluating soil in large areas in which large numbers of samples are available. Similarly, Schwartz et al. (2012) obtained more accurate results at a lower cost using VIS-NIR-SWIR spectroscopy to monitor oil contamination in soils, compared to traditional methods of analytical chemistry.

The quantification of soil attributes, especially SOC, with a sensor is achieved by building prediction models using the analytical and spectral values. Table 1 provides a compilation of several studies that built predictive models using different multivariate methods and spectral pre-processing (treatments) techniques in local, regional, national and global scale spectral libraries. In these studies, the definition of the scale is related to the territorial coverage of the soil samples collected, which is variable. Another aspect is that the accuracy of the SOC predictions varies among studies.

There are several possible explanations for this variation in the accuracy of prediction models. First, the type of pre-processing applied to the spectral data and the multivariate prediction methods used to calibrate the models may differ. Studies have shown that different spectral pre-processing transformations result in different predictive performances for SOC (Dunn et al., 2002; Vasques et al., 2008; Knox et al., 2015; Dotto et al., 2017). Regarding multivariate methods, Vasques et al. (2008) obtained better results using partial least squares regression (PLSR) compared with multiple linear regression (MLR). Nonetheless, Vašát et al. (2017) observed the opposite. In studies by Sequeira et al. (2014) and Knox et al. (2015), the random forest (RF) showed advantages in comparison to PLSR when the SOC values were not normally distributed. Viscarra Rossel and Behrens (2010) verified that support vector machines (SVMs) generated more accurate models when compared to PLSR and RF. Lucà et al. (2017) observed a better predictive capacity of SVM relative to PLSR and principal component regression (PCR) and differences according to the number of samples in the calibration dataset. In contrast, Dotto et al. (2017) did not observe any difference between the PLSR and SVM methods for SOC prediction.

Second, the variation in the accuracy of the models is related to the intrinsic complexity of the soil, which is a natural resource. Mineralogy, particle size, moisture and SOC content interfere with the absorption of electromagnetic radiation, causing variations in reflectance (Stevens et al., 2013; Wight et al., 2016). Therefore, the heterogeneity of the sample set might lead to a bias in the estimation of attributes in studies that use the VIS-NIR-SWIR spectroscopic technique (Bellon-Maurel and McBratney, 2011). In the review by Stenberg et al. (2010), the authors mentioned the existence of a positive correlation between the prediction error and the standard deviation of the soil properties under investigation. This result has been corroborated by Demattê et al. (2016), who used spectral libraries on regional and local scales. The authors reported that the low variability in some attributes was a limiting factor in building more accurate models, and the presence of soils with high variations in components affects the prediction of soil chemical properties in local scale studies. To circumvent this problem, researchers have focused on developing strategies to improve the accuracy of predictions, such as combining information from local and national spectral data (Gogé et al., 2014).

The variability in SOC predictions found in the literature demonstrated the need for additional studies using spectral data focusing on a local scale, which can be incorporated into spectral libraries of greater scale in the future. In addition, there is a lack of scientific knowledge concerning the type of spectral pre-processing, multivariate methods and strategies to build predictive models for SOC in local spectral libraries of subtropical soils in Brazil. Considering the hypothesis that the accuracy of the models can be increased through the use of different preprocessing techniques, multivariate methods and stratification of the spectral library in groups, the objectives of this study were to i) characterize and identify differences among spectra of the soil classes, land uses and sampling layers in relation to spectral behavior, ii) assess different pre-processing techniques and multivariate methods to propose SOC prediction models based on the spectral data and, iii) evaluate the performance of SOC prediction models generated from the stratification of a local library of soil VIS-NIR-SWIR spectra.

2. Materials and methods

2.1. Description of the study area

The study was conducted in an area of 940 ha located in the city of Giruá, State of Rio Grande do Sul (RS), Brazil (Fig. 1a). The area belongs to the physiographic region named *Planalto do RS*, with altitudes varying between 339 and 426 m. The predominant climate is Cfa, according to Köppen's classification, which is characterized as a humid subtropical climate without a defined dry season (Alvares et al., 2013). The relief goes from gently undulated to undulated, with more pronounced slopes near drainages. Hills with slopes ranging from 3 to 10% are common. The geology of the area is derived from the *Formação Serra Geral*. Soil classes that occur in the area are Rhodic Ferralsol (Dystric, Clayic) – FR and Dystric Gleysol (Clayic, Humic) – GL (IUSS/WRB, 2015). Predominant land uses in the area are native forest (NFo), native fields (NFi) and crops in a no-tillage system (CTS).

2.2. Field sampling and soil analysis

Soil samples were collected at 261 locations at depths ranging from 0.00 to 0.05 (C1), 0.05 to 0.15 (C2) and 0.15 to 0.30 m (C3) (C1 + C2 + C3 = 783 soil samples), and 29 locations at depths ranging from 0.30 to 0.60 (C4) and 0.60 to 1.00 m (C5) (C4 + C5 = 58 soil samples) according to the specifications of the GlobalSoilMap consortium (Arrouays et al., 2014), resulting in a total of 841 soil samples (Fig. 1b). At each location, three samples were collected to create a single composite sample for the laboratory analysis. The sampling locations in the field were defined in two ways: (1) an irregular sampling grid with 400 points was created, with the objective of representing all the environmental variability in the area; (2) from the 400 points, 261 were selected based on tacit knowledge of the Pedologist regarding land use, soil class and geomorphological units. The objective of this sampling methodology was to capture the entire variability of the area and incorporate the tacit knowledge of the Pedologist.

In the laboratory, all the samples were air-dried, ground, sieved (2-mm mesh), and submitted to chemical analysis in triplicate. The SOC content was determined via wet combustion with external heating