Revista Brasileira de Ciência do Solo

Division - Soil in Space and Time | Commission - Pedometric

A Regional Legacy Soil Dataset for Prediction of Sand and Clay Content with Vis-Nir-Swir, in Southern Brazil

Elisângela Benedet Silva[1]*, **Élvio Giasson**[2], **André Carnieletto Dotto**[3], **Alexandre ten Caten**[4], **José Alexandre Melo Demattê**[3], **Ivan Luiz Zilli Bacic**[1] **and Milton da Veiga**[5]

[1] Empresa de Pesquisa Agropecuária e Extensão Rural de Santa Catarina, Florianópolis, Santa Catarina, Brasil.
[2] Universidade Federal do Rio Grande do Sul, Departamento de Ciência do Solo, Porto Alegre, Rio Grande do Sul, Brasil.
[3] Universidade de São Paulo, Escola Superior de Agricultura "Luiz de Queiroz", Departamento de Ciência do Solo, Piracicaba, São Paulo, Brasil.
[4] Universidade Federal de Santa Catarina, Departamento de Ciência Veterinária e Biologia, Curitibanos, Santa Catarina, Brasil.
[5] Universidade do Oeste de Santa Catatina, Curso de Agronomia, Campos Novos, Santa Catarina, Brasil.

ABSTRACT: The success of soil prediction by VIS-NIR-SWIR spectroscopy has led to considerable investment in large soil spectral libraries. The aims of this study were 1) to develop a soil VIS-NIR-SWIR spectroscopy approach using legacy soil samples to improve spectral soil information in a regional scale; (2) to compare six spectral preprocessing techniques; and (3) to compare the performance of linear and non-linear multivariate models for prediction of sand and clay content. A total of 1,534 legacy soil samples, stored by Epagri, were collected from agricultural areas in 2009 on a regional scale, covering 260 municipalities of Santa Catarina. Six spectral preprocessing techniques were applied and compared with reflectance spectra (control treatment) in the development of sand and clay prediction models. Five multivariate regression models, Support Vector Machines, Gaussian Process Regression, Cubist, Random Forest, and Partial Least Square Regression were compared. The scatter-corrective preprocessing groups produced similar or better performance than spectral-derivatives. In addition, preprocessing spectra prior to regression analysis does not improve sand prediction, since reflectance spectra achieved the best performance using Cubist, SVM, and PLS models. In general, clay content presented better prediction accuracy than sand content. The best multivariate model to predict sand and clay content from soil VIS-NIR-SWIR spectra was Cubist. The best Cubist performance was achieved combined with reflectance spectra ($R^2 = 0.73$; root mean square error = 10.60 %; ratio of the performance to the interquartile range = 2.36) and MSC ($R^2 = 0.83$; root mean square error = 7.29 %; ratio of the performance to the interquartile range = 3.70) for sand and clay content, respectively. Considering the mean RMSE values of the validation set, the predictive ability of the multivariate models decreased in the following order: Cubist>PLS>RF>GPR>SVM for both properties. The predictive ability of VIS-NIR-SWIR reflectance spectroscopy achieved in this study for sand and clay content using legacy soil data and heterogeneous samples confirmed the potential of the spectroscopy approach.

Keywords: soil spectral library, multivariate models, preprocessing techniques, Santa Catarina.

*** Corresponding author:**
E-mail: elisbenedetsilva@gmail.com

Received: November 28, 2018

Approved: May 06, 2019

How to cite: Silva EB, Giasson E, Dotto AC, ten Caten A, Demattê JAM, Bacic ILZ, Veiga M. A regional legacy soil dataset for prediction of sand and clay content with VIS-NIR-SWIR, in Southern Brazil. Rev Bras Cienc Solo. 2019;43:e0180174.
https://doi.org/10.1590/18069657rbcs20180174

INTRODUCTION

Reflectance spectroscopy in the visible, near and shortwave infrared (VIS-NIR-SWIR) regions has been proposed as a rapid, accurate, and cost-effective model to predict chemical, physical, and mineralogical properties using laboratory, field, and airborne hyperspectral sensors (Vasques et al., 2008; Demattê et al., 2016a; Nouri et al., 2017; Poggio et al., 2017; Viscarra Rossel et al., 2017; Dotto et al., 2018). Soil VIS-NIR-SWIR spectra are non-specific and include weak, wide, and overlapping absorption bands directly linked to soil composition, whereby moisture, particle size, organic matter, and mineralogy of the clay fraction and iron oxides influence spectral behavior (Stenberg et al., 2010).

Different preprocessing techniques have been applied to transform soil spectra, removing noise from the multiple scattering effect, highlighting specific features of the spectra, eliminating redundant information, and preparing the soil spectra for spectral modeling (Rinnan et al., 2009). As reported by Buddenbaum and Steffens (2012), these techniques represent an important step in the multivariate approach and include several algorithms such as smoothing, normalization, scatter-correction, continuum removal, and derivatives. Some studies have reported improvements in the performance of prediction models (Vasques et al., 2008; Nawar et al., 2016; Dotto et al., 2017), while others found similar or better results with no spectral preprocessing (Sawut et al., 2014; Viscarra Rossel and Webster, 2012). The type and amount of required preprocessing techniques are site-specific (Stenberg et al., 2010) and, with large datasets, the effects of preprocessing steps are not clear (Engel et al., 2013).

Several multivariate models based on VIS-NIR-SWIR have been applied to processing soil spectra in order to mathematically extract meaningful information from individual spectrum to accurately predict chemical and physical soil properties, such as organic carbon/matter, pH, total nitrogen, soil moisture, and cation exchange capacity, among others (Morellos et al., 2016; Demattê et al., 2017; Dotto et al., 2018; Xu et al., 2018). The capacity to predict sand, silt, and clay has also been demonstrated in previous studies (Vendrame et al., 2012; Demattê et al., 2016b; Lacerda et al., 2016; Nawar et al., 2016; Dotto et al., 2017; Santana et al., 2018), but none of them in a regional soil legacy spectral library of subtropical soils in Brazil. Among the multivariate model, the partial least square regression (PLS) is the most common multivariate model used (Dotto et al., 2018), given its simplicity and robustness (Viscarra Rossel et al., 2006; Vasques et al., 2008; Lacerda et al., 2016). However, other studies have established that nonlinear data-mining models such as Support Vector Machines (SVM), Gaussian Process Regression (GPR), and Random Forest (RF) can outperform PLS when used to build predictive models from reflectance spectra (Terra et al., 2015; Nawar et al., 2016; Dotto et al., 2017; Santana et al., 2018). In addition to these models, another data-mining tool based on Cubist regression-rules has been introduced into the spectroscopy approach to predicting soil properties (Minasny and Mcbratney, 2008; Viscarra Rossel and Webster, 2012; Morellos et al., 2016; Viscarra Rossel et al., 2016; Zeng et al., 2017; Sorenson et al., 2018). Minasny and McBratney (2008), Morellos et al. (2016), and Sorenson et al. (2018) used Cubist to build predictive models of soil properties, including clay content, total carbon, total nitrogen, moisture content, and cation exchange capacity, and reported that Cubist provided better results than those provided by PLS.

The success of soil prediction by VIS-NIR-SWIR has led to considerable investment in large soil spectral libraries (Shepherd and Walsh, 2002; Brown et al., 2006; Viscarra Rossel and Webster, 2012). Soil information stored by universities, research centers, and government agencies, among others, could provide an opportunity to enlarge spectral libraries for data-poor regions, new challenges with regard to the reliability of such data and brings understanding to improve future site sampling (Nocita et al., 2015; Viscarra Rossel et al., 2016). While several studies have been done to build predictive models for soil properties based on local, regional, and national spectral libraries in Brazil (Bellinaso et al., 2010; Ramirez-Lopes et al., 2013; Araújo et al., 2014; Demattê et al., 2016b; Lacerda et al., 2016), the application of soil reflectance spectroscopy has not been reported on a regional scale in South Brazil, especially in the state of Santa Catarina (SC). In this state, existing studies using soil spectral libraries are limited to local scale with low variability (Dotto et al., 2017, 2018). This research aims to fill this gap and to further the use of reflectance spectroscopy for assessing sand and clay content using a legacy soil dataset in subtropical soils based on a regional spectral library.

Given the high variability of our regional scale legacy soil dataset, the hypothesis stated is that the performance of the prediction models will rely on a combination of preprocessing, multivariate models, and soil property being predicted. The main objectives of this study were: (1) to explore the potential of a legacy soil dataset with large range variability of sand and clay content on a regional scale, to predict soil properties by a systematic VIS-NIR-SWIR spectroscopy approach; (2) to compare six spectral preprocessing techniques in the development of sand and clay content models; and (3) to compare the performance of linear and non-linear multivariate models for prediction of sand and clay content.

MATERIALS AND METHODS

Study area and soil spectral library

The study area covers 260 municipalities (about 90 %) of the state of Santa Catarina (Figure 1). The state is characterized by its diversity in climate, vegetation, geology, relief, and soil. Santa Catarina (SC) has two climate types according to the Köppen classification system (Figure 1); super humid and mesothermal (Cfa) and quite humid and mesothermal (Cfb). The remaining original vegetation includes areas of Rain Forest type and five major subtypes, these being Dense Rain Forest, Araucaria Forest, Alpine Grassland, Deciduous Forest, and Coastal Vegetation (Klein, 1978). The geology consists of granitoids, charnockitics, gneisses, and granites in Eastern SC, the Gondwana Plateau

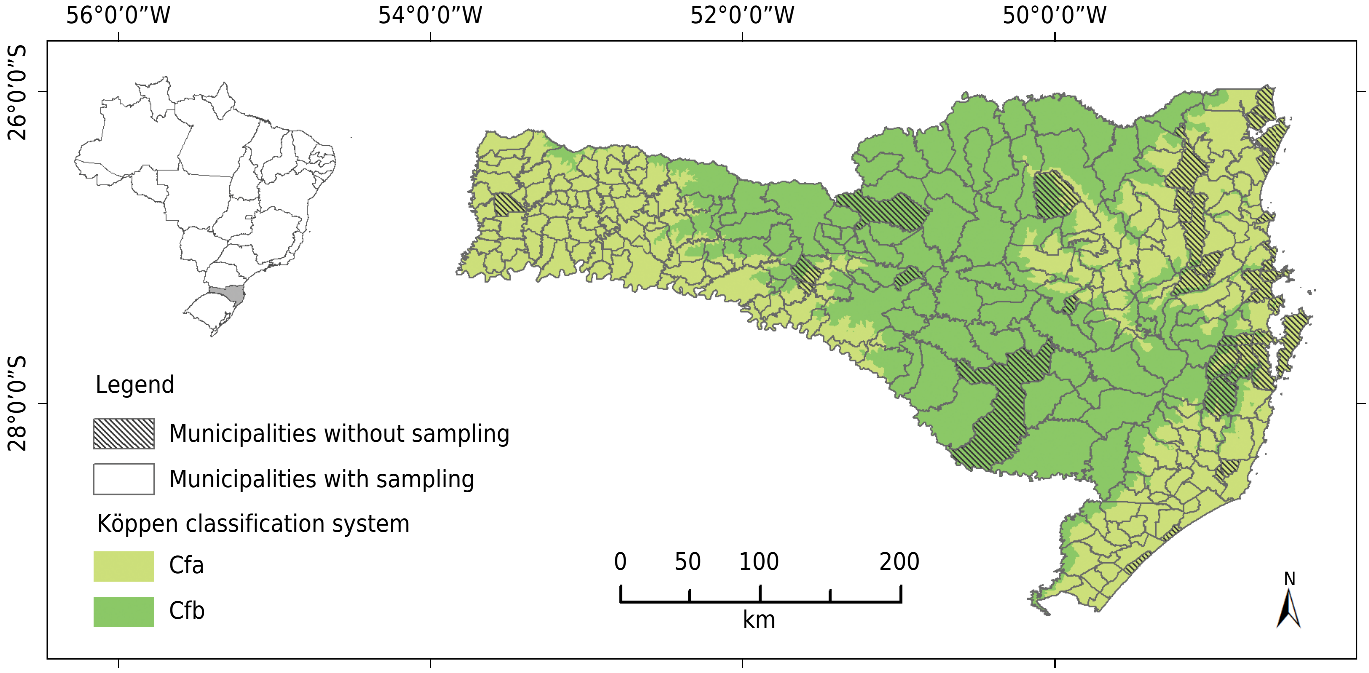
56°0'0"W
54°0'0"W
52°0'0"W
50°0'0"W
26°0'0"S
28°0'0"S
Legend
Municipalities without sampling
Municipalities with sampling
Köppen classification system
Cfa
Cfb
0
50
100
200
km
N

Figure 1. Sampling of soil data from municipalities of Santa Catarina State.

and a basalt plateau in Western SC, with a predominance of basic volcanic rocks and quartz sandstones, with siliceous and argillaceous intercalations (Silva and Bortoluzzi, 1987). Soils are diverse and the predominant soil order, according to Brazilian Soil Classification System (Santos et al., 2013) and the IUSS Working Group WRB (2014), are *Cambissolos* (Inceptisols, 46.0 % of SC area), *Neossolos* (Entisols, 18.5 %), *Nitossolos* (Ultisols, Oxisols, 13.8 %), *Argissolos* (Ultisols, 7.7 %), *Latossolos* (Oxisols, 6.1 %), *Gleissolos* (Aquents, 4.2 %) (Embrapa, 2004).

The soil dataset provided by the Agricultural Research and Rural Extension Corporation of Santa Catarina (Epagri - *Empresa de Pesquisa Agropecuária e Extensão Rural de Santa Catarina*) was used to develop the soil spectral library. A total of 1,534 samples were collected from agricultural areas in 2009 on a regional scale, from within the 0.00-0.50 m soil layer and described in Veiga et al. (2012). Samples were air-dried, ground, sieved to 2 mm, and stored. The sand (0.05-2 mm) and clay (<0.002 mm) content were determined in 2009 according to the pipette standard method (Teixeira et al., 2017). The silt fraction was not considered in this study.

Spectral measurements

The spectral reflectance of soil samples was obtained using a FieldSpec 3 spectroradiometer (Analytical Spectral Devices Inc., Boulder, CO, USA) in the VIS-NIR-SWIR range (350-2500 nm), following standard laboratory procedure of the Brazilian Soil Spectral Library (Romero et al., 2018). The samples were placed in a petri dish and shaken to ensure a smooth surface for spectrum acquisition. The light source was two halogen (50 W) bulbs with the beam non-collimated to the target plane, positioned at a distance of 35 cm from the sample with a zenithal angle of 30°. The spectral sensor captured the light through a fiber-optic cable connected to the sensor, placed vertically within 8 cm of the sample, where the reflected light in an area of approximately 2 cm^2 at the center of the sample was measured. As a reference standard, a white Spectralon® was used at the beginning of the measurements and after every 20 readings. Each spectrum measurement was the result of the average of 50 sensor readings. A total of three scans were collected from each sample, rotating the petri dish by 90° for each scan, and these were then averaged to obtain a representative spectrum.

Spectral preprocessing

In this study, six spectral preprocessing techniques were applied and compared in the development of sand and clay prediction models. The techniques were divided into groups of scatter-correction and spectral-derivatives. The first group included: (i) multiplicative scatter correction (MSC), which removes additive and/or multiplicative signal effects (Martens and Naes, 1992); (ii) detrending (DET), to reduce the effect of particle size and additionally remove the linear, or curvilinear, trend of each spectrum (Barnes et al., 1989); and (iii) normalizations by range (NBR), to get all data to approximately the same scale; (iv) continuum removed reflectance (CRR), which removes the continuous features of the spectra and isolates specific absorption features present in the spectrum to minimize noise (Clark and Roush, 1984). The second group is spectral derivatives, represented by (v) Savitzky–Golay first derivative using a first-order polynomial with a search window of 11 nm (FDSG) and (vi) Savitzky–Golay second derivative using a second order polynomial with a search window of 11 nm (SDSG).

The preprocessing techniques (Figure 2) were selected because they have been shown to improve the performance of spectroscopy models and have different effects on model prediction. They were applied to soil reflectance curves in the range of 350-2,500 nm. Reflectance spectra (RAW) with no spectral preprocessing applied was used as “control treatment”. All preprocessing techniques were carried out using R (R Development Core Team, 2017) by applying *prospectr* (Ramirez-Lopes et al., 2013), *pls* (Mevik et al.,

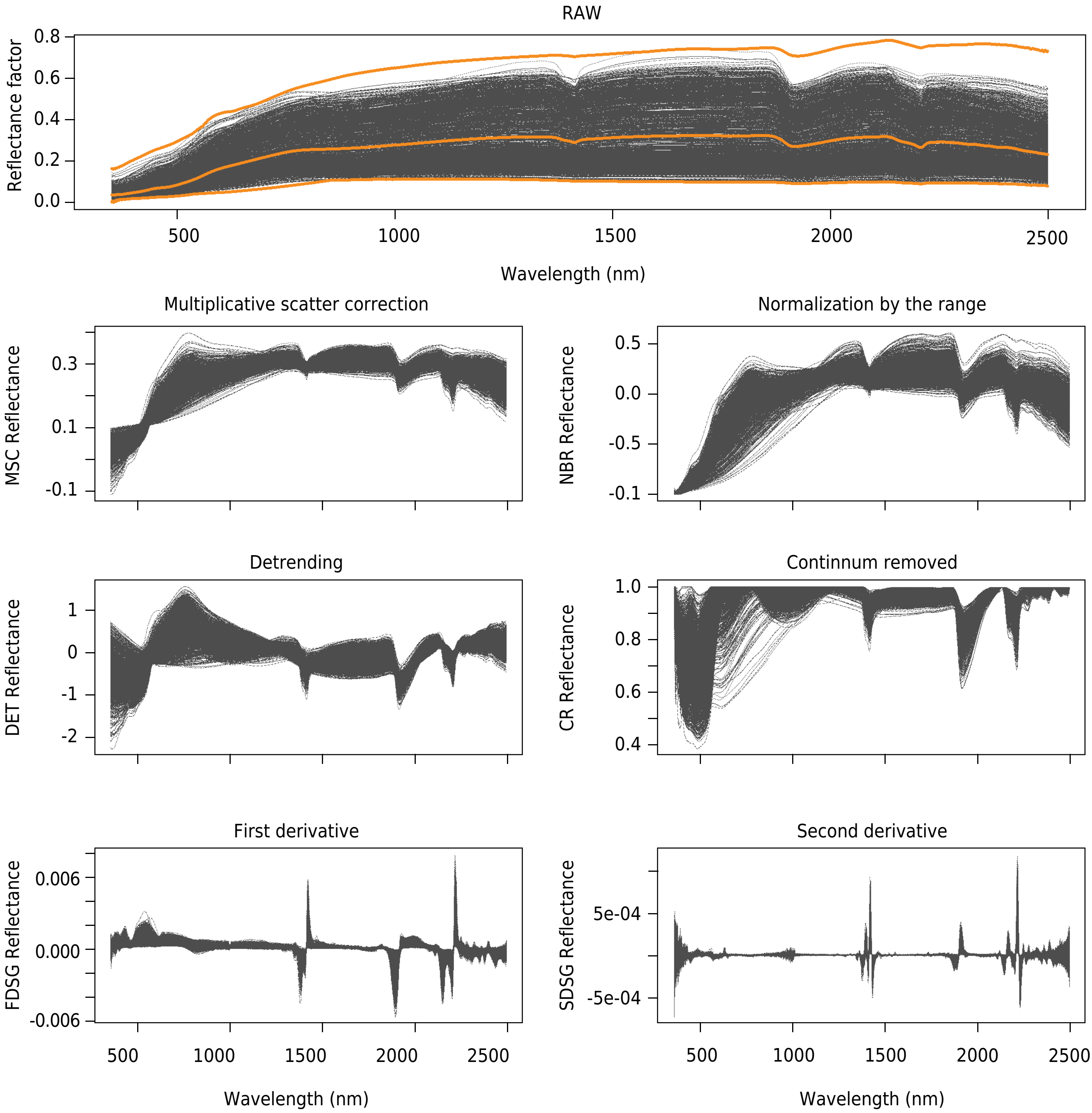

Figure 2. VIS-NIR-SWIR reflectance spectra and the preprocessing techniques of the spectral curves for all soil samples. The highlighted raw spectra correspond to the mean, the lowest, and the highest reflectance spectra of the soil samples.

2016), and *clusterSim* (Dudek, 2017) packages. The spectral preprocessing techniques presented here are discussed in more detail in Rinnan et al. (2009) and Buddenbaum and Steffens (2012).

To better compare the preprocessing techniques, including RAW spectra, the SK test of R^2 and RMSE mean values of the validation set was carried out. The Scott-Knott (SK) test (Scott and Knott, 1974) was used to compare the average values of R^2 and RMSE between different models and preprocessing techniques to verify significant differences between them. The SK performs a hierarchical cluster analysis approach used to partition treatment into distinct homogeneous groups by minimizing variation within groups and maximizing variation between groups (Scott and Knott, 1974). The cluster procedure begins with the whole group of observed mean effects and then divides and keeps dividing subgroups in such a way that the intersection of any of the two formed groups remains empty (Jelihovschi et al., 2014). It was applied using the *ScottKnott* package (Jelihovschi et al., 2014).

Multivariate models

To compare the performance of the proposed multivariate regression models using a regional spectral library, we compared to two supervised learning algorithms with linear kernel function, Support Vector Machines (SVM) and Gaussian Process Regression (GPR), two tree-based models, Cubist and Random Forest (RF), and one of the most common linear model used in the spectroscopy approach, Partial Least Square Regression (PLS). The regression process was implemented based on the measured reflectance spectra (RAW and six spectral preprocessing techniques) and the measured values of sand and clay content using the training set. The predictive models were assessed for each soil property using the independent validation set. Only the best predictive model, laboratory-measured versus VIS-NIR-SWIR-predicted values of sand and clay content, will be plotted. The modeling was performed using several packages in R (R Development Core Team, 2017) and the parameters of each model were manually optimized to generate the best possible fit between the variables and outputs.

The multivariate model used were Cubist, regression-rules model (Holmes et al., 1999), Random Forest (RF) as an ensemble learning model (Breiman, 2001), and Support Vector Machine (SVM) as a machine technique based on statistical learning theory (Vapnik, 1995).

Gaussian Process Regression (GPR) as a probabilistic, non-parametric Bayesian approach, and PLSR (Wold et al., 2001). The RMSE was used in this study to identify the number of latent factors and *leave-one-out cross-validation* was used for the model training set to verify prediction performance (Boos, 2003) and to prevent over or under-fitting the data in the training step. From the total dataset (n = 1,534), 75 % were separated at random into the training set (n = 1,151), to create the regression models, while the remaining 25 % (n = 383, validation set) were used to independently validate the models. To check the reliability of splitting of each subset, the Levene's test and Student's t-test were applied to verify the equality of variances and means, respectively. The coefficient of determination (R^2), root mean square error (RMSE), and the ratio of performance to the interquartile range (RPIQ) were used to assess the performance of sand and clay content prediction models, using equations (1), (2), and (3), respectively. The R^2 provides the proportion of the variance explained by the model. The RMSE provides the accuracy of predictions, giving the standard deviation of the model prediction error in the same units as the attributes. The RPIQ was used instead of the ratio of performance deviation (RPD) because it is based on quartile range and better represents the spread of the population for skewed distributions (Bellon-Maurel et al., 2010). The highest R^2 (Equation 1) and RPIQ (Equation 2) values and the lowest RMSE (Equation 3) value of the validation set were used to determine the selection of the best spectral preprocessing and multivariate regression models.

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \qquad \text{Eq. 1}$$

where $\hat{y}_i$ is the value predicted by the model; y_i is the measured value; $\bar{y}_i$ is the average value; and N is the number of samples.

$$RPIQ = \frac{IQ}{RMSE} \qquad \text{Eq. 2}$$

where IQ is the interquartile distance (IQ = Q3-Q1) of the observed values, which accounts for 50 % of the population around the median.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - \bar{y}_i)^2} \qquad \text{Eq. 3}$$

where N is the number of samples used in the prediction; and $\hat{y}_i$ and y_i are the values of predicted and measured soil properties, respectively.

RESULTS

The samples under study presented a wide variation with sand and clay contents (Table 1), indicating great variability in terms of particle size distribution. The independent training and validation sets showed Levene's test *p-value* of 0.609 and of 0.175 for sand and clay content values in the training and validation sets, respectively. According to Student's t test sand (*p-value* = 0.179) and clay (*p-value* = 0.435), did not show a significant difference at a 5 % significance level, for the training and validation sets, respectively.

The effect of six preprocessing techniques fluctuated between models (Tables 2 and 3), so it is difficult to reach a clear conclusion as to whether the differences between the average values of R^2 and RMSE among preprocessing techniques are significant. It was observed that there was no statistical difference between the mean values of R^2 and RMSE for RAW spectra, NBR, MSC, DET, CRR, and FDSG for sand and clay prediction models (Figure 3a). In other words, on average, RAW spectra, NBR, MSC, DET, CRR, and FDSG preprocessing techniques had an equal effect on model performance to quantify sand and clay content, and they clearly perform better than SDSG preprocessing. Thus, RAW spectra and MSC preprocessing are the best strategies for sand and clay content, as they consistently simplify the models. The worst results for both R^2 and RMSE were found with SDSG preprocessing (Figure 3b). There is significant variability across SDSG results. The poorest result was achieved by SDSG-SVM, which presented the lowest R^2 and the highest RMSE (0.19 and 27.94 % for sand and 0.28 and 23.13 % for clay).

The predictive model performance of sand content presented R^2, RMSE, and RPIQ values ranging from 0.19 to 0.73, 10.60 to 27.9 %, and 0.9 to 2.4, respectively (Table 2). The predictive model performance for clay content presented R^2, RMSE, and RPIQ values ranging from 0.28 to 0.83, 7.3 to 23.1 %, and 1.2 to 3.7, respectively. For sand, the SK test showed that there was no statistical difference between R^2 and RMSE mean values of the five models applied, using a 5 % significant level. For clay, the SK test showed

Table 1. Descriptive statistics of soil properties for dataset, training set and validation set

	Dataset (100 %)		Training set (75 %)		Validation set (25 %)	
	Sand	Clay	Sand	Clay	Sand	Clay
Observations	1534	1534	1151	1151	383	383
Minimum	1.00	0.00	1.00	0.00	3.00	0.00
Maximum	99.00	77.00	98.00	76.00	99.00	77.00
Mean	28.85	38.25	28.30	38.45	30.53	37.63
Median	25.00	37.00	25.00	38.00	27.00	36.00
Std error of mean	0.50	0.40	0.55	0.50	1.09	0.91
Skewness	1.10	0.11	1.08	0.11	1.04	0.05
Kustosis	1.10	-0.80	1.17	-0.75	0.78	-0.83
CV	67	45	67	44	70	47

Table 2. Performance of sand predictive models from five multivariate methods with the corresponding spectral preprocessing techniques

Soil property	Method	Preprocessing	Validation set		
			R^{2*}	RMSE (%)	RPIQ
Sand	Cubist	RAW	0.73	10.6	2.4
		MSC	0.70	11.0	2.2
		NBR	0.67	11.7	2.1
		DET	0.65	12.0	2.1
		FDSG	0.64	12.0	2.1
		CRR	0.63	12.4	2.0
		SDSG	0.62	12.4	2.0
	RF	FDSG	0.68	11.6	2.2
		SDSG	0.63	12.8	2.0
		DET	0.61	12.8	2.0
		CRR	0.61	12.9	1.9
		MSC	0.61	12.8	1.9
		NBR	0.60	13.1	1.9
		RAW	0.57	13.4	1.9
	SVM	RAW	0.67	11.6	2.2
		NBR	0.64	12.1	2.1
		CRR	0.61	12.5	2.0
		MSC	0.60	12.8	2.0
		DET	0.58	13.0	1.9
		FDSG	0.58	13.4	1.9
		SDSG	0.19	27.9	0.9
	PLS	RAW	0.67	11.6	2.1
		FDSG	0.65	11.9	2.1
		NBR	0.63	12.2	2.0
		MSC	0.60	12.7	2.0
		CRR	0.60	12.7	2.0
		DET	0.59	12.9	1.9
		SDSG	0.54	13.7	1.8
	GPR	NBR	0.65	12.0	2.1
		RAW	0.64	12.1	2.1
		CRR	0.62	12.5	2.0
		MSC	0.61	12.5	2.0
		FDSG	0.61	12.8	1.9
		DET	0.58	13.0	1.9
		SDSG	0.30	20.8	1.2
Minimum			0.19	10.6	0.9
Maximum			0.73	27.9	2.4
Mean			0.60	13.1	2.0
Standard Deviation			0.10	3.0	0.26

* Sorted by ascending order of R^2 (coefficient of determination for validation set).

different results for R^2 and RMSE between the Cubist (R^2 = 0.79, RMSE = 8.06 %) and the remaining multivariate models, PLS (R^2 = 0.69, RMSE = 10.08 %), RF (R^2 = 0.68, RMSE = 10.15 %), GPR (R^2 = 0.65, RMSE = 11.07 %), and SVM (R^2 = 0.62, RMSE = 12.13 %) presented a statistical difference (Figure 4).

Table 3. Performance of clay predictive models from five multivariate methods with the corresponding spectral preprocessing techniques

Soil property	Method	Preprocessing	Validation set		
			R^{2*}	RMSE (%)	RPIQ
Clay	Cubist	MSC	0.83	7.3	3.7
		RAW	0.83	7.3	3.7
		DET	0.83	7.4	3.7
		NBR	0.82	7.5	3.6
		CRR	0.79	8.1	3.3
		FDSG	0.74	9.1	3.0
		SDSG	0.69	9.8	2.8
	SVM	NBR	0.77	8.7	3.1
		MSC	0.76	8.7	3.1
		DET	0.73	9.4	2.9
		RAW	0.69	9.9	2.7
		FDSG	0.63	11.6	2.3
		CRR	0.45	13.5	2.0
		SDSG	0.28	23.1	1.2
	GPR	NBR	0.76	8.8	3.1
		MSC	0.75	9.0	3.0
		DET	0.71	9.6	2.8
		RAW	0.67	10.3	2.6
		FDSG	0.67	10.5	2.6
		CRR	0.57	11.9	2.3
		SDSG	0.42	17.3	1.6
	RF	FDSG	0.76	8.8	3.0
		DET	0.70	10.0	2.7
		CRR	0.68	10.2	2.6
		MSC	0.68	10.3	2.6
		SDSG	0.69	10.3	2.6
		NBR	0.65	10.6	2.5
		RAW	0.63	10.9	2.5
	PLS	NBR	0.75	8.9	3.0
		MSC	0.73	9.3	2.9
		CRR	0.72	9.5	2.8
		DET	0.70	9.8	2.7
		RAW	0.70	10.1	2.7
		FDSG	0.67	10.5	2.6
		SDSG	0.55	12.4	2.2
Minimum			0.28	7.3	1.2
Maximum			0.83	23.1	3.7
Mean			0.69	10.3	2.8
Standard Deviation			0.12	2.9	0.5

* Sorted by ascending order of R^2 (coefficient of determination for validation set).

DISCUSSION

Effects of the preprocessing techniques on modeling

In general, model performances (Tables 2 and 3) decreased as the noise resulting from preprocessing increased (from RAW to SDSG). The derivatives emphasize noise in the data more distinctly than the other methods (Buddenbaum and Steffens, 2012). Even

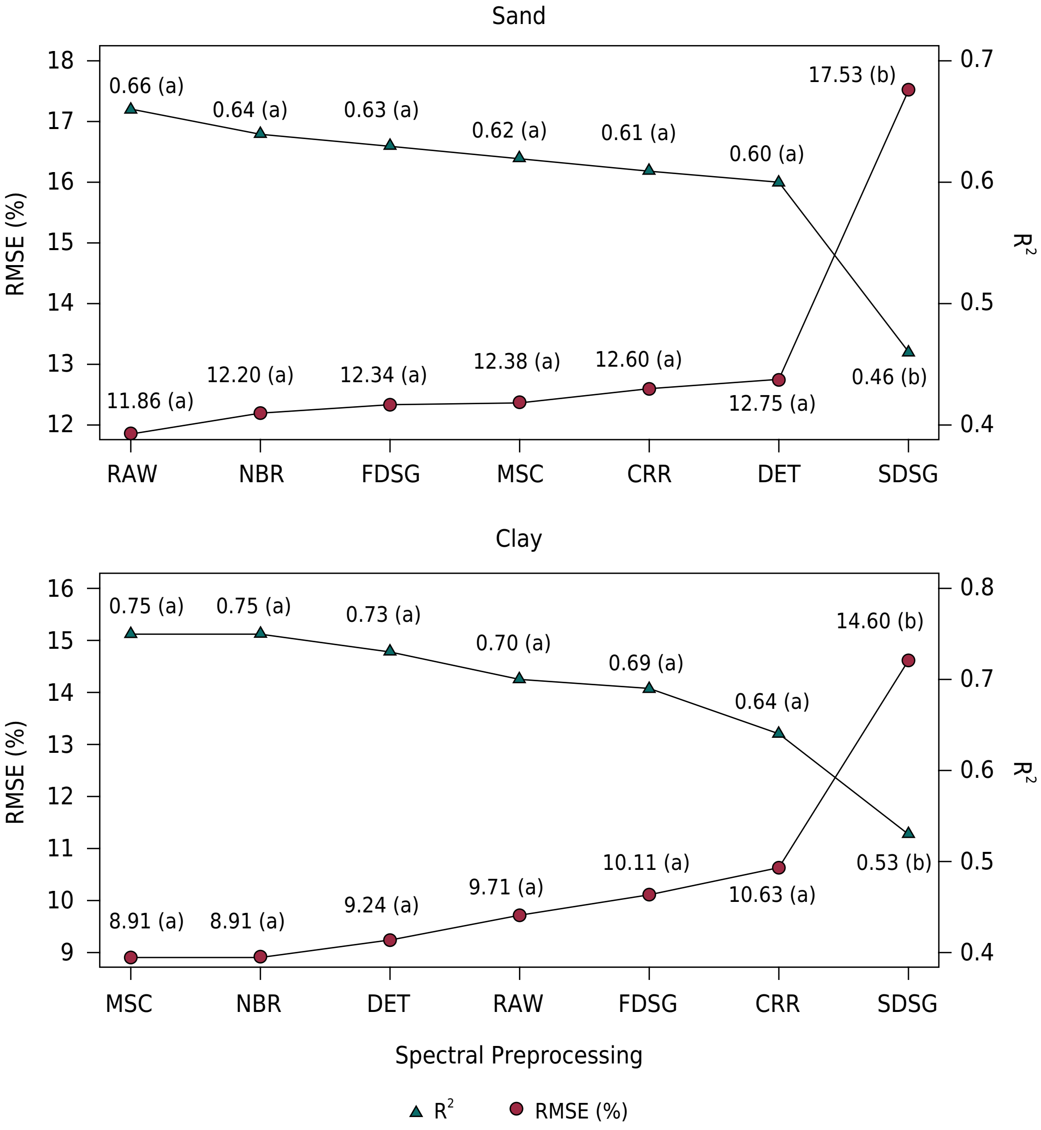

Figure 3. The mean values of RMSE and R^2 for each preprocessing technique for sand and clay content. The letters in parentheses represent the results of the Scott Knott test (significance level α=0.05).

though extremely flat structures can be evaluated with spectral derivatives (FDSG and SDSG), this also tends to increase spectral noise (García-Sánchez et al., 2017), especially by second spectral derivatives (Buddenbaum and Steffens, 2012). For sand, RF (with FDSG and SDSG) did not appear to be sensitive to enhanced noise in the spectrum (Breiman, 2001), revealing its capability to better handle derivative transformation (Dotto et al., 2018). However, SDSG decreased model performance of the remaining multivariate models for the two soil properties studied. These results with spectral derivatives and ensemble-learning algorithms (RF) are in agreement with Vasques et al. (2008), Pinheiro et al. (2017), Dotto et al. (2018), and Santana et al. (2018).

Considering the performance of RAW, preprocessing the spectra before regression analysis did not improve sand prediction, with the exception when used with RF models. Therefore, spectral reflectance values without preprocessing (RAW spectra) were sufficient to obtain highly accurate models, and our results show that there is no need to perform any preprocessing technique on the spectra to generate better prediction models for sand in the VIS-NIR-SWIR region. The results achieved in the current study are divergent from Franceschini et al. (2013), which found high performance of the sand model (R^2 = 0.87) with applying spectral preprocessing. Considering the GPR model, only a slight benefit

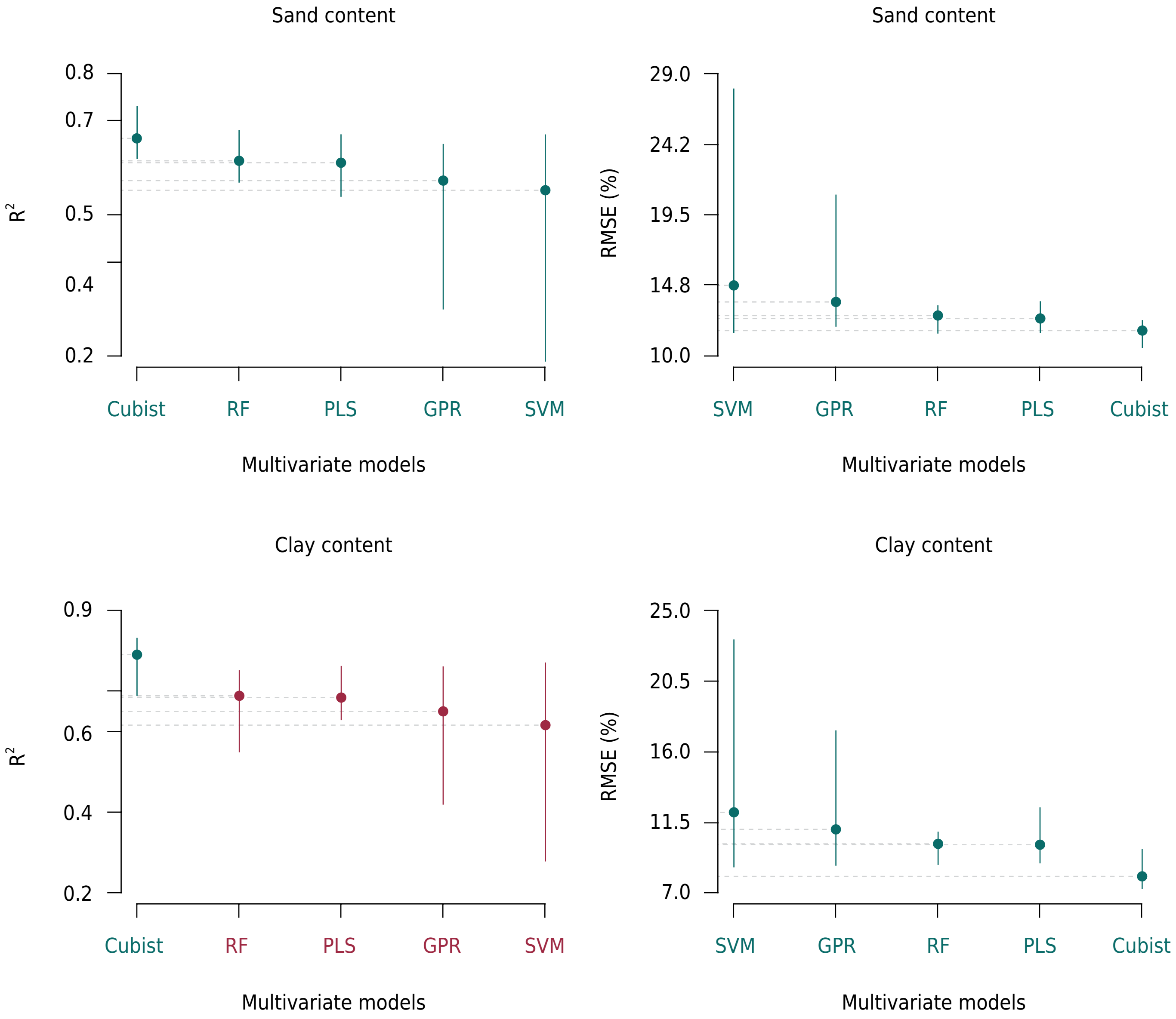

Figure 4. Statistical difference between multivariate methods resulted by Scott Knott test (significance level of 10 %). The mean, maximum, and minimum values of R^2 and RMSE for each method applied for sand and clay content.

was found using preprocessing (NBR) on the spectra. This finding is in agreement with Duda et al. (2017), who reported no significant improvement in results with first derivative preprocessing compared with RAW spectra using the SVM model. They combined two proximal sensor approaches relative to a single sensor, to compare the efficacy in determining soil properties, sand and clay content, among others, in a catena scale in Eastern Europe. Sawut et al. (2014) reported a small influence of preprocessing on spectral analysis for the prediction of sand content in a thermal infrared region. All these studies worked with a smaller number of samples and range variability of the sand content than ours.

Preprocessing results for FDSG derivatives are in agreement with those found by Bilgili et al. (2010), Pinheiro et al. (2017), and Duda et al. (2017), who reported different R^2 values (0.84, 0.62, and 0.25, respectively). Spectral preprocessing may emphasize the feature sought in the spectra and several authors have noted its benefit (Vasques et al., 2008; Nawar et al., 2016; Dotto et al., 2018). Sand fractions, tends to have quartz as the dominant mineral (Demattê et al., 2007) which has no diagnostic spectral features in the VIS-NIR-SWIR ranges, high values of reflectance intensity (albedo), and its reflectance spectrum is largely unvarying (Hunt and Salisbury, 1970; Clark, 1999; Ramirez-Lopes et al., 2013; Wight et al., 2016). Preprocessing techniques are designed for baseline corrections, so their effect is minimal when the baseline variance is small (Buddenbaum and Steffens, 2012). In general, sandy soils exhibit similar spectral behavior of quartz, and this may explain why no preprocessing technique was very helpful in improving sand model performance in the present study.

For clay, MSC and NBR achieved the best performance with the SVM, GPR, and PLS models. These preprocessing techniques are normalization procedures commonly used to compensate for baseline shift and multiplicative effects in the spectral data, which are induced by physical effects such as particle size (Martens and Naes, 1992; Rinnan et al., 2009; García-Sánchez et al., 2017). The MSC, which attempts to eliminate the effects of the spectrum by linearizing each spectrum by the average spectrum of the sample, is the most popular normalization technique (Martens and Naes, 1992). In NBR, each spectrum is divided by the range. In this study, MSC (R^2 = 0.70, RMSE = 11.0 %, RPIQ = 2.4) exhibited slightly better performance than NBR (R^2 = 0.67, RMSE = 11.7 %, RPIQ = 2.1), when combined with Cubist model (Table 2), albeit with the same performance as RAW spectra (control treatment). The NBR produced the best model result for the SVM, GPR, and PLS models.

The CRR presented inferior performance than expected, given that other studies commonly report this technique as an effective VIS-NIR-SWIR data preprocessing technique (Lagacherie et al., 2008; Nawar et al., 2016; Dotto et al., 2017, 2018). In general, the scatter-corrective group gives similar, or better, performance than spectral-derivatives for sand and clay models. These results are in agreement with Dotto et al. (2017), who also reported better performance of the models with scatter-corrective preprocessing techniques compared to spectral-derivatives to predict soil organic carbon using a local spectral library from SC. All multivariate models applied in the present study achieved different performance with two spectral preprocessing groups. For clay content, the accuracy of the prediction models appears to depend on spectral preprocessing, except for Cubist model that achieved the same performance of prediction with and without spectral preprocessing technique. For sand, this was not so evident. The SVM seems to be more sensitive to spectral preprocessing applications, since R^2 dropped from 0.67 to 0.19 and from 0.77 to 0.28 for sand and clay content, respectively. The RMSE showed the inverse trend.

Effects of the multivariate models

Considering the performance of the models, Lacerda et al. (2016) developed a PLS prediction model to quantify soil texture from 3,750 soil samples using the topossequence model from three areas of São Paulo State, Brazil. These authors found predictions values in the validation set for sand (R^2 = 0.96, RMSE = 137.98 g kg^{-1}) and clay (R^2 = 0.93, RMSE = 82.50 g kg^{-1}) content. The R^2 and RMSE values are higher than those found in this study. The lowest R^2 found with the SC dataset can be explained in terms of the high heterogeneity of soil samples collected throughout the state. The soil samples used in this study are legacy soil samples and were sampled for another purpose than this spectroscopy study. Further, they were collected from the depth soil layer (0.00-0.50 m) showing great variability of the sand and clay content into that depth in soils with texture gradient. On the other hand, legacy soil samples bring new challenges and lead the techniques to limit.

Higher performance was achieved by Terra et al. (2015) in the VIS-NIR-SWIR region for clay (R^2 = 0.85, RMSE = 96.50 g kg^{-1}) and sand content (R^2 = 0.85, RMSE = 25.22 g kg^{-1}), using 1,259 soil samples from four Brazilian states. In Nawar et al. (2016) the R^2 and RMSE for clay content in the validation set ranged from 0.52 to 0.79 % and 11.35 to 7.75 %, respectively, using different multivariate models and preprocessing techniques based on a spectral library with limited soil samples (n = 102) from Northern Sinai, Egypt.

For sand, the best results were obtained by Cubist (Table 2) based on RAW spectra (R^2 = 0.73, RMSE = 10.60 %, RPIQ = 2.36) and MSC preprocessing techniques (R^2 = 0.70, RMSE = 11.09 %, RPIQ = 2.25), followed by lower performances using RF with FDSG (R^2 = 0.68, RMSE = 11.56 %, RPIQ = 2.16), SVM-RAW (R^2 = 0.67, RMSE = 11.60 %, RPIQ = 2.16), and PLS-RAW (R^2 = 0.67, RMSE = 11.62 %, RPIQ = 2.15). The RF outperformed the SVM, PLS, and GPR models only with FDSG, whereas with all remaining spectral preprocessing showed similar performance (Table 2). Compared to published results using Cubist model (R^2 = 0.50) (e.g., Zeng et al., 2017), RF (R^2 = 0.82) (e.g., Santana et al., 2018), SVM (R^2 = 0.25-0.90) (Viscarra Rossel and Behrens, 2010; Terra et al., 2015; Dotto et al., 2017), and PLSR (R^2 = 0.33-0.96) (Wetterlind et al., 2008; Sawut et al., 2014; Terra et al., 2015; Lacerda et al., 2016; Dotto et al., 2017; Pinheiro et al., 2017; Conforti et al., 2018) the results of the present study showed good quantitative predictions. However, all these studies used different soils, different methodologies, multivariate models, and preprocessing techniques, with different population sizes, and VIS-NIR-SWIR or MID-IF in some situations.

The scatter plot (Figure 5) of laboratory-measured versus VIS-NIR-SWIR-predicted values of sand and clay content based on Cubist model using the validation set showed quite low dispersion, with most of the values distributed close to the 1:1 line (red line), with small slope and intercept values, indicating good fit (Viscarra Rossel and Webster, 2012).

The Cubist model showed the best performance in comparison with the other models for the prediction of clay content (Table 3). The R^2 values of the validation set ranged from 0.69 to 0.83, whereas RMSE ranged from 9.79 to 7.29 %. This confirms the superior performance of Cubist in predicting soil properties, as stated in the literature (Minasny and Mcbratney, 2008; Viscarra Rossel and Webster, 2012; Stevens et al., 2013; Morellos et al., 2016; Viscarra Rossel et al., 2016). This result is in line with results

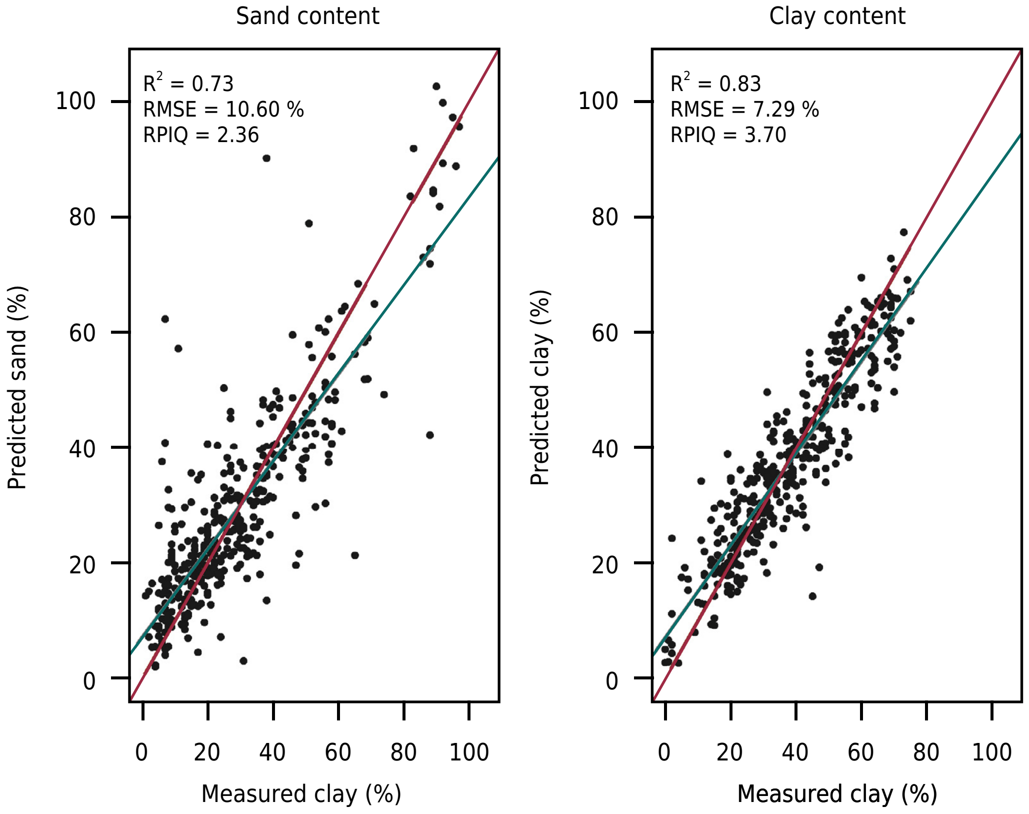

Figure 5. Assessment set of laboratory-measured versus VIS-NIR-SWIR-predicted values using Cubist (with RAW spectra) for sand and (with MSC spectra) for clay. Red line indicates 1:1 line.

reported by Viscarra Rossel and Webster (2012), who used Cubist to predict 24 soil properties, including sand, silt, and clay content, using a large soil dataset (n = 21,493) from all the states in Australia and found RMSE = 12.00 % and RMSE = 8.49 % for sand and clay content, respectively. They concluded that the rule-based model predicts sand and clay content well, working effectively with large and diverse datasets. A study presented by Minasny and McBratney (2008) showed that Cubist produced the best fit model (R^2 = 0.92) and lowest error (RMSE = 7.18 %) when compared to PLS and another data-mining model, Treenet, using mid-infrared (2500-25000 nm) spectra of soil samples from Australia.

Scatter-correction (MSC and NBR) spectral preprocessing performed well and provided the best results for Cubist-MSC (R^2 = 0.83, RMSE = 7.29 %, RPIQ = 3.70), SVM-NBR (R^2 = 0.77, RMSE = 8.69 %, RPIQ = 3.11), GPR-NBR (R^2 = 0.76, RMSE = 8.80 %, RPIQ = 3.07), and PLS-NBR (R^2 = 0.75, RMSE = 8.94 %, RPIQ = 3.02), respectively. These results were consistent with previous studies estimating clay content based on VIS-NIR-SWIR regions (R^2 = 0.62-0.93) (Minasny and Mcbratney, 2008; Viscarra Rossel and Behrens, 2010; Viscarra Rossel and Webster, 2012; Ramirez-Lopes et al., 2013; Araújo et al., 2014; Terra et al., 2015; Lacerda et al., 2016; Dotto et al., 2017; Lucà et al., 2017; Pinheiro et al., 2017; Santana et al., 2018). Similar performance was achieved with RF-FDSG (R^2 = 0.76, RMSE = 8.85 %, RPIQ = 3.05). Machine learning algorithms (Cubist, SVM, GPR, and RF) outperformed the PLS approach (Table 3). This better performance may be explained by the inclusion of non-linear relationship between clay content and spectra, interaction effects of the regression task, as well as linear combinations of variables (Kovačević et al., 2010; Gomez et al., 2016). In the study presented by Dotto et al. (2017), SVM and PLS models were applied to the prediction of soil organic carbon (SOC), sand, silt, and clay content using VIS-NIR-SWIR ranges. A local scale dataset of 299 soil samples from the central region of SC was used and statistical differences between the RMSE mean values of the SVM (RMSE = 7.68 %) and PLS (RMSE = 8.58 %) models were found, whereby SVM produced the best fitting model (R^2_{val} = 0.62) and the lowest error ($RMSE_{val}$ = 6.84 %) for clay content estimation. For SOC, sand, and silt, they did not find any statistical differences between the two multivariate models. On the other hand, Santana et al. (2018) compared RF and PLS to assess sand and clay content and found a significant difference between the results of the two models, RF ($RMSE_{val}$ = 7.61 %) and PLS ($RMSE_{val}$ = 8.82 %) for clay content, with RF proving to be the better approach. These authors used 641 soil samples from several regions of Brazil.

The PLS is the most linear common multivariate model for quantitative spectroscopy analysis in soil. This model is based on the decomposition of spectral data into latent variables that capture most of the variance existing in the spectrum, and linear models are then created using the scores of the most correlated features (Morellos et al., 2016). However, in PLS, non-linear relationships can only be modeled in a limited way, and the model is a linear function of all wavenumbers, whereas in regression-rule models like Cubist, these non-linearities can be efficiently modeled using a set of comprehensible linear equations (Minasny and Mcbratney, 2008).

In the SK test, Cubist (R^2 = 0.66, RMSE = 11.73 %), RF (R^2 = 0.62, RMSE = 12.76 %), PLS (R^2 = 0.61, RMSE = 12.54 %), GPR (R^2 = 0.57, RMSE = 13.68 %), and SVM (R^2 = 0.55, RMSE = 14.77 %) presented the same performance in sand prediction (Figure 4). However, when comparing R^2 and RMSE performance for each model (Figure 4), it was clear that Cubist achieved the best performance, followed by regular performance by the PLS, RF, and GPS models, with the poorest performance being found for SVM, although it remained acceptable for sand prediction.

For clay, the SK test results showed that the models were divided into two groups (blue and dark red color groups, Figure 4), of which Cubist presented the best performance. However, for the mean RMSE values, there was no statistical difference in the SK test (α = 10 %). The prediction quality decreased from Cubist to SVM, with mean RMSE values increasing from 8.06 to 12.13 %, suggesting different performances between the models. These results again demonstrated the better performance of the Cubist model in comparison to the most common algorithms used in spectroscopy analysis.

CONCLUSIONS

The performance of preprocessing techniques fluctuated between models. In addition, spectra preprocessing before regression analysis does not improve sand prediction. Scatter-corrective preprocessing groups (MSC, NBR, DET, and CRR) produced similar or better performance than spectral-derivatives (FDSG and SDSG). Spectral-derivatives only showed better results with RF models for both attributes.

The best multivariate model to predict sand and clay content from soil VIS-NIR-SWIR spectra was Cubist.

The predictive ability of the multivariate models decreased in the following order: Cubist>PLS>RF>GPR>SVM for sand and clay.

These legacy soil data can produce reliable information, and it can be used to populate a soil database as input to soil monitoring, as a primary reference to establish standards of the spectral behavior of soils in the Santa Catarina state.

ACKNOWLEDGMENTS

The authors would like to thank the Coordination for the Improvement of Higher Education Personnel (CAPES), Brazilian National Council for Scientific and Technological Development (CNPq), and Agricultural Research and Rural Extension Corporation of Santa Catarina (EPAGRI) for the financial support and FAPESP (Project No. 2014/22262-0) for the equipment support and the Geotechnologies in Soil Science Group, GEOSS (http://esalqgeocis.wixsite.com/english). The authors would also like to thank the journal editor and reviewers for their useful comments.

AUTHOR CONTRIBUTIONS

Conceptualization: Elisângela Benedet da Silva and Alexandre ten Caten.

Methodology: Elisângela Benedet da Silva and André Carnieletto Dotto.

Formal Analysis: Elisângela Benedet da Silva.

Investigation: Elisângela Benedet da Silva.

Resources: Elvio Giasson and Alexandre Mello Demattê.

Data Curation: Elisângela Benedet da Silva, Alexandre Mello Demattê, and Milton da Veiga.

Writing – Original Draft: Elisângela Benedet da Silva.

Writing – Review &amp; Editing: Elisângela Benedet da Silva, Alexandre ten Caten, André Carnieletto Dotto, Milton da Veiga, Elvio Giasson, Ivan Luiz Zilli Bacic, and Alexandre Mello Demattê.

Visualization: Elisângela Benedet da Silva.

Supervision: Elvio Giasson and Ivan Luiz Zilli Bacic.

Project Administration: Elisângela Benedet da Silva.

Funding Acquisition: Ivan Luiz Zilli Bacic, Elvio Giasson, and Alexandre Mello Dematté.

REFERENCES

Araújo SR, Wetterlind J, Demattê JAM, Stenberg B. Improving the prediction performance of a large tropical vis-NIR spectroscopic soil library from Brazil by clustering into smaller subsets or use of data mining calibration techniques. Eur J Soil Sci. 2014;65:718-29. https://doi.org/10.1111/ejss.12165

Barnes RJ, Dhanoa MS, Lister SJ. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Appl Spectrosc. 1989;43:772-7. https://doi.org/10.1366/0003702894202201

Bellinaso H, Demattê JAM, Romeiro SA. Soil spectral library and its use in soil classification. Rev Bras Cienc Solo. 2010;34:861-70. https://doi.org/10.1590/S0100-06832010000300027

Bellon-Maurel V, Fernandez-Ahumada E, Palagos B, Roger JM, McBratney A. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. TrAC - Trend Anal Chem. 2010;29:1073-81. https://doi.org/10.1016/j.trac.2010.05.006

Bilgili AV, van Es HM, Akbas F, Durak A, Hively WD. Visible-near infrared reflectance spectroscopy for assessment of soil properties in a semi-arid area of Turkey. J Arid Environ. 2010;74:229-38. https://doi.org/10.1016/j.jaridenv.2009.08.011

Boos DD. Introduction to the bootstrap world. Stat Sci. 2003;18:168-74.

Breiman L. Random forests. Mach Learning. 2001;45:5-32.

Brown DJ, Shepherd KD, Walsh MG, Mays MD, Reinsch TG. Global soil characterization with VNIR diffuse reflectance spectroscopy. Geoderma. 2006;132:273-90. https://doi.org/10.1016/j.geoderma.2005.04.025

Buddenbaum H, Steffens M. The effects of spectral pretreatments on chemometric analyses of soil profiles using laboratory imaging spectroscopy. Appl Environ Soil Sci. 2012;2012:274903. https://doi.org/10.1155/2012/274903

Clark RN. Spectroscopy of rocks and minerals, and principles of spectroscopy. In: Rencz AN, editor. Remote sensing for the earth sciences: manual of remote sensing. 3rd ed. New York: John Wiley; 1999. v. 3. p. 3-58.

Clark RN, Roush TL. Reflectance spectroscopy: quantitative analysis techniques for remote sensing applications. JGR Solid Earth. 1984;89:6329-40. https://doi.org/10.1029/JB089iB07p06329

Conforti M, Matteucci G, Buttafuoco G. Using laboratory Vis-NIR spectroscopy for monitoring some forest soil properties. J Soils Sediments. 2018;18:1009-19. https://doi.org/10.1007/s11368-017-1766-5

Demattê JAM, Alves MR, Terra FS, Bosquilia RWD, Fongaro CT, Barros PPS. Is it possible to classify topsoil texture using a sensor located 800 km away from the surface? Rev Bras Cienc Solo. 2016a;40:e0150335. https://doi.org/10.1590/18069657rbcs20150335

Demattê JAM, Bellinaso H, Araújo SR, Rizzo R, Souza AB. Spectral regionalization of tropical soils in the estimation of soil attributes. Rev Cienc Agron. 2016b;47:589-98. https://doi.org/10.5935/1806-6690.20160071

Demattê JAM, Nanni MR, Formaggio AR, Epiphanio JCN. Spectral reflectance for the mineralogical evaluation of Brazilian low clay activity soils. Int J Remote Sens. 2007;28:4537-59. https://doi.org/10.1080/01431160701250408

Demattê JAM, Ramirez-Lopez L, Marques KPP, Rodella AA. Chemometric soil analysis on the determination of specific bands for the detection of magnesium and potassium by spectroscopy. Geoderma. 2017;288:8-22. https://doi.org/10.1016/j.geoderma.2016.11.013

Dotto AC, Dalmolin RSD, Grunwald S, ten Caten A, Pereira Filho W. Two preprocessing techniques to reduce model covariables in soil property predictions by Vis-NIR spectroscopy. Soil Till Res. 2017;172:59-68. https://doi.org/10.1016/j.still.2017.05.008

Dotto AC, Dalmolin RSD, ten Caten A, Grunwald S. A systematic study on the application of scatter-corrective and spectral- derivative preprocessing for multivariate prediction of soil organic carbon by Vis-NIR spectra. Geoderma. 2018;314:262-74. https://doi.org/10.1016/j.geoderma.2017.11.006

Duda BM, Weindorf DC, Chakraborty S, Li B, Man T, Paulette L, Deb S. Soil characterization across catenas via advanced proximal sensors. Geoderma. 2017;298:78-91. https://doi.org/10.1016/j.geoderma.2017.03.017

Dudek MW. ClusterSim: searching for optimal clustering procedure for a data set [internet]. Jelenia Góra: Wrocław University of Economics; 2017. Available from: http://keii.ue.wroc.pl/clusterSim.

Empresa Brasileira de Pesquisa Agropecuária - Embrapa. Solos do estado de Santa Catarina. Rio de Janeiro: Embrapa Solos; 2004. (Boletim de Pesquisa e Desenvolvimento, 46).

Engel J, Gerretzen J, Szymańska E, Jansen JJ, Downey G, Blanchet L, Buydens LMC. Breaking with trends in pre-processing? TrAC - Trend Anal Chem. 2013;50:96-106. https://doi.org/10.1016/j.trac.2013.04.015

Franceschini MHD, Demattê JAM, Sato MV, Vicente LE, Grego CR. Abordagens semiquantitativa e quantitativa na avaliação da textura do solo por espectroscopia de reflectância bidirecional no VIS-NIR-SWIR. Pesq Agropec Bras. 2013;48:1569-82. https://doi.org/10.1590/S0100-204X2013001200006

García-Sánchez F, Galvez-Solo L, Martínez-Nicolás JJ, Muelas-Domingo R, Nieves M. Using near-infrared spectroscopy in agricultural systems. In: Kyprianidis KG, Skvaril J. Developments in near-infrared spectroscopy. London: IntechOpen; 2017. p. 97-127.

Gomez C, Gholizadeh A, Borůvka L, Lagacherie P. Using legacy data for correction of soil surface clay content predicted from VNIR/SWIR hyperspectral airborne images. Geoderma. 2016;276:84-92. https://doi.org/10.1016/j.geoderma.2016.04.019

Holmes G, Hall M, Prank E. Generating rule sets from model trees. In: Foo N, editor. Proceedings 12th Australian Joint Conference on Artificial Intelligence, AI’99: Advanced topics in artificial intelligence. December; 1999; Sydney. Berlin: Springer; 1999. p. 1-12.

Hunt GR, Salisbury JW. Visible and near-infrared spectra of minerals and rocks: I silicate minerals. Modern Geology. 1970;1:283-300.

IUSS Working Group WRB. World reference base for soil resources 2014, update 2015: International soil classification system for naming soils and creating legends for soil maps. Rome: Food and Agriculture Organization of the United Nations; 2015. (World Soil Resources Reports, 106).

Jelihovschi EG, Faria JC, Allaman IB. ScottKnott: a package for performing the Scott-Knott clustering algorithm in R. Tend Mat Apl Comput. 2014;15:3-17. https://doi.org/10.5540/tema.2014.015.01.0003

Klein RM. Mapa fitogeográfico do estado de Santa Catarina. Itajaí: Ioesc; 1978.

Kovačević M, Bajat B, Gajić B. Soil type classification and estimation of soil properties using support vector machines. Geoderma. 2010;154:340-7. https://doi.org/10.1016/j.geoderma.2009.11.005

Lacerda MPC, Demattê JAM, Sato MV, Fongaro CT, Gallo BC, Souza AB. Tropical texture determination by proximal sensing using a regional spectral library and its relationship with soil classification. Remote Sens. 2016;8:701. https://doi.org/10.3390/rs8090701

Lagacherie P, Baret F, Feret J-B, Madeira Netto J, Robbez-Masson JM. Estimation of soil clay and calcium carbonate using laboratory, field and airborne hyperspectral measurements. Remote Sens Environ. 2008;112:825-35. https://doi.org/10.1016/j.rse.2007.06.014

Lucà F, Conforti M, Castrignanò A, Matteucci G, Buttafuoco G. Effect of calibration set size on prediction at local scale of soil carbon by Vis-NIR spectroscopy. Geoderma. 2017;288:175-83. https://doi.org/10.1016/j.geoderma.2016.11.015

Martens H, Naes T. Multivariate calibration. In: Kowalski BR, editor. Chemometrics: mathematics and statistics in chemistry. Dordrecht: Springer-Science; 1992. p. 147-56.

Mevik B-H, Wehrens R, Liland KH. R Package: PLS: partial least squares and principal component regression [internet]; 2016. Available from: https://cran.r-project.org/package=pls

Minasny B, Mcbratney AB. Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. Chemometr Intell Lab. 2008;94:72-9. https://doi.org/10.1016/j.chemolab.2008.06.003

Morellos A, Pantazi X-E, Moshou D, Alexandridis T, Whetton R, Tziotzios G, Wiebensohn J, Bill R, Mouazen AM. Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. Biosyst Eng. 2016;152:104-16. https://doi.org/10.1016/j.biosystemseng.2016.04.018

Nawar S, Buddenbaum H, Hill J, Kozak J, Mouazen AM. Estimating the soil clay content and organic matter by means of different calibration methods of VIS-NIR diffuse reflectance spectroscopy. Soil Till Res. 2016;155:510-22. https://doi.org/10.1016/j.still.2015.07.021

Nocita M, Stevens A, van Wesemael B, Aitkenhead M, Bachmann M, Barthès B, Ben Dor E, Brown DJ, Clairotte M, Csorba A, Dardenne P, Demattê JAM, Genot V, Guerrero C, Knadel M, Montanarella L, Noon C, Ramirez-Lopes L, Robertson J, Sakai H, Soriano-Disla JM, Shepherd KD, Stenberg B, Towett EK, Vargas R, Wetterlind J. Soil spectroscopy : an alternative to wet chemistry for soil monitoring. Adv Agron. 2015;132:139-59. https://doi.org/10.1016/bs.agron.2015.02.002

Nouri M, Gomez C, Gorretta N, Roger JM. Clay content mapping from airborne hyperspectral Vis-NIR data by transferring a laboratory regression model. Geoderma. 2017;298:54-66. https://doi.org/10.1016/j.geoderma.2017.03.011

Pinheiro EFM, Ceddia MB, Clingensmith CM, Grunwald S, Vasques GM. Prediction of soil physical and chemical properties by visible and near-infrared diffuse reflectance spectroscopy in the central Amazon. Remote Sens. 2017;9:293. https://doi.org/10.3390/rs9040293

Poggio M, Brown DJ, Bricklemyer RS. Comparison of Vis-NIR on *in situ*, intact core and dried, sieved soil to estimate clay content at field to regional scales. 2017;68:434-48. https://doi.org/10.1111/ejss.12434

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria; 2017. Available from: http://www.R-project.org/.

Ramirez-Lopes L, Behrens T, Schmidt K, Stevens A, Demattê JAM, Scholten T. The spectrum-based learner: a new local approach for modeling soil VIS-NIR spectra of complex datasets. Geoderma. 2013;195-196:268-79. https://doi.org/10.1016/j.geoderma.2012.12.014

Rinnan Å, van den Berg F, Engelsen SB. Review of the most common pre-processing techniques for near-infrared spectra. TrAC - Trends Anal Chem. 2009;28:1201-22. https://doi.org/10.1016/j.trac.2009.07.007

Romero DJ, Ben-Dor E, Demattê JAM, Souza AB, Vicente LE, Tavares TR, Martello M, Strabeli TF, Barros PPS, Fiorio PR, Gallo BC, Sato MV, Eitelwein MT. Internal soil standard method for the Brazilian soil spectral library: Performance and proximate analysis. Geoderma. 2018;312:95-103. https://doi.org/10.1016/j.geoderma.2017.09.014

Santana FB, Souza AM, Poppi RJ. Visible and near infrared spectroscopy coupled to random forest to quantify some soil quality parameters. Spectrochim Acta A. 2018;191:454-62. https://doi.org/10.1016/j.saa.2017.10.052

Sawut M, Ghulam A, Tiyip T, Zhang Y-j, Ding J-l, Zhang F, Maimaitiyiming M. Estimating soil sand content using thermal infrared spectra in arid lands. Int J Appl Earth Obs Geoinf. 2014;33:203-10. https://doi.org/10.1016/j.jag.2014.05.010

Scott AJ, Knott M. A Cluster analysis method for grouping means in the analysis of variance. Biometrics. 1974;30:507-12. https://doi.org/10.2307/2529204

Shepherd KD, Walsh MG. Development of reflectance spectral libraries for characterization of soil properties. Soil Sci Soc Am J. 2002;66:988-98. https://doi.org/10.2136/sssaj2002.0988

Silva LC, Bortoluzzi CA. Mapa geológico do Estado de Santa Catarina, escala 1:500.000: texto explicativo. Florianópolis: DNPM/Secretaria de Ciência, Tecnologia, Minas e Energia/ Coordenadoria de Recursos Minerais; 1987.

Sorenson PT, Quideau SA, Rivard B. High resolution measurement of soil organic carbon and total nitrogen with laboratory imaging spectroscopy. Geoderma. 2018;315:170-7. https://doi.org/10.1016/j.geoderma.2017.11.032

Stenberg B, Rossel RAV, Mouazen AM, Wetterlind J. Visible and near infrared spectroscopy in soil science. Adv Agron. 2010;107:163-215. https://doi.org/10.1016/S0065-2113(10)07005-7

Stevens A, Nocita M, Tóth G, Montanarella L, van Wesemael B. Prediction of soil organic carbon at the European scale by visible and near infrared reflectance spectroscopy. PLoS ONE. 2013;8:e66409. https://doi.org/10.1371/journal.pone.0066409

Teixeira PC, Donagemma GK, Fontana A, Teixeira WG. Manual de métodos de análise de solo. 3. ed. rev e ampl. Brasília, DF: Embrapa; 2017.

Terra FS, Demattê JAM, Rossel RAV. Spectral libraries for quantitative analyses of tropical Brazilian soils: Comparing vis–NIR and mid-IR reflectance data. Geoderma. 2015;255-256:81-93. http://dx.doi.org/10.1016/j.geoderma.2015.04.017

Vapnik VN. The nature of statistical learning theory. 2nd ed. New York: Springer; 1995.

Vasques GM, Grunwald S, Sickman JO. Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra. Geoderma. 2008;146:14-25. https://doi.org/10.1016/j.geoderma.2008.04.007

Veiga M, Santos O, Hammes LA, Pandolfo C. Distribuição espacial dos teores de argila, silte e areia na camada superficial do solo em Santa Catarina. Rev Agropec Catarinense. 2012;25:63-8.

Vendrame PRS, Marchão RL, Brunet D, Becquer T. The potential of NIR spectroscopy to predict soil texture and mineralogy in Cerrado Latosols. Eur J Soil Sci. 2012;63:743-53. https://doi.org/10.1111/j.1365-2389.2012.01483.x

Viscarra Rossel RA, Behrens T. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma. 2010;158:46-54. https://doi.org/10.1016/j.geoderma.2009.12.025

Viscarra Rossel RA, Behrens T, Ben-Dor E, Brown DJ, Demattê JAM, Shepherd KD, Shi Z, Stenberg B, Stevens A, Adamchuk V, Aïchi H, Barthès BG, Bartholomeus HM, Bayer AD, Bernoux M, Böttcher K, Brodský L, Du CW, Chappell A, Fouad Y, Genot V, Gomez C, Grunwald S, Gubler A, Guerrero C, Hedley CB, Knadel M, Morrás HJM, Nocita M, Ramirez-Lopez L, Roudier P, Campos EMR, Sanborn P, Sellitto VM, Sudduth KA, Rawlins BG, Walter C, Winowiecki LA, Hong SY, Ji W. A global spectral library to characterize the world’s soil. Earth-Sci Rev. 2016;155:198-230. https://doi.org/10.1016/j.earscirev.2016.01.012

Viscarra Rossel RA, Lobsey CR, Sharman C, Flick P, McLachlan G. Novel proximal sensing for monitoring soil organic C stocks and condition. Environ Sci Technol. 2017;51:5630-41. https://doi.org/10.1021/acs.est.7b00889

Viscarra Rossel RA, Walvoort DJJ, McBratney AB, Janik LJ, Skjemstad JO. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma. 2006;131:59-75. https://doi.org/10.1016/j.geoderma.2005.03.007

Viscarra Rossel RA, Webster R. Predicting soil properties from the Australian soil visible-near infrared spectroscopic database. Eur J Soil Sci. 2012;63:848-60. https://doi.org/10.1111/j.1365-2389.2012.01495.x

Wetterlind J, Stenberg B, Jonsson A. Near infrared reflectance spectroscopy compared with soil clay and organic matter content for estimating within-field variation in N uptake in cereals. Plant Soil. 2008;302:317-27. https://doi.org/10.1007/s11104-007-9489-9

Wight JP, Ashworth AJ, Allen FL. Organic substrate, clay type, texture, and water influence on NIR carbon measurements. Geoderma. 2016;261:36-43. https://doi.org/10.1016/j.geoderma.2015.06.021

Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. Chemometr Intell Lab. 2001;58:109-30. https://doi.org/10.1016/S0169-7439(01)00155-1

Xu S, Zhao Y, Wang M, Shi X. Comparison of multivariate methods for estimating selected soil properties from intact soil cores of paddy fields by Vis-NIR spectroscopy. Geoderma. 2018;310:29-43. https://doi.org/10.1016/j.geoderma.2017.09.013

Zeng R, Rossiter DG, Yang F, Li D-C, Zhao Y-G, Zhang G-L. How accurately can soil classes be allocated based on spectrally predicted physio-chemical properties? Geoderma. 2017;303:78-84. https://doi.org/10.1016/j.geoderma.2017.05.011