Soil Security 23 (2026) 100236

Contents lists available at ScienceDirect

Soil Security

journal homepage: www.sciencedirect.com/journal/soil-security

Spectrally-based soil carbon models to support the National Forest Inventory of Rio de Janeiro, Brazil

Gustavo M. Vasques [a,*], Levi B. Luz [b], Fabiano C. Balieiro [a], Monise A. F. Magalhães [c], Telmo B. Silveira Filho [c], Marcelo T. Andrade [a]

[a] Embrapa Solos, Rua Jardim Botânico 1024, Rio de Janeiro, RJ 22460-000, Brazil
[b] Universidade Federal Fluminense, Outeiro de São João Batista s/n, Campus do Valonguinho, Niterói, RJ 24020-150, Brazil
[c] Secretaria de Estado do Ambiente e Sustentabilidade do Rio de Janeiro. Av. Venezuela, 110, Rio de Janeiro, RJ 20081-312, Brazil

ARTICLE INFO

Keywords:
Soil spectral library
Machine learning
Chemometrics
Spectral modeling

ABSTRACT

Methods for assessing soil carbon must be fast and accurate to support soil health and security evaluation, measurement, reporting and verification of carbon stocks. Visible-near infrared (Vis-NIR) spectroscopy minimizes the costs and time required for assessing soil carbon. Soil samples were obtained at 189 sites from the National Forest Inventory (NFI) of Rio de Janeiro state (~43,782 km^2), Brazil, between 2013 and 2016, at 0–20 and 30–50 cm, with a total of 373 recovered samples. The objective was to compare different preprocessing transformations of spectral curves, and calibration methods to predict soil carbon contents from soil Vis-NIR spectral curves in NFI samples from Rio de Janeiro state. Soil carbon contents were measured by dry combustion in a CHNS elemental analyzer, and soil spectral curves were obtained in the laboratory. Cubist combined with log(1/reflectance) preprocessing emerged as the optimal combination to predict soil carbon content (root mean square error of cross-validation of 5.1 g kg^{-1}), whereas elastic net obtained good soil carbon content predictions consistently in both cross-validation and external validation. Partial least squares regression, random forest, support vector machines, and model ensemble produced poorer results. The results agree with previous studies comparing calibration methods for soil carbon content prediction, and stress the importance of preprocessing soil spectral curves, as well as testing different methods to produce robust results. Soil Vis-NIR spectroscopy can be used to assess soil carbon contents in Rio de Janeiro, supporting expedited and accurate soil carbon stock, and stock change assessments in future phases of the NFI.

1. Introduction

Soils are a crucial resource for producing food, fiber, and energy, essential for supporting human life and sustaining terrestrial ecosystems. They play a key role in the biogeochemical cycles of major nutrient elements. Carbon contributes to soil quality by regulating nutrients and toxic substances, storing water, stabilizing soil aggregates and structure, and controlling microbial activity, ultimately influencing biodiversity and ecosystem sustainability (Banwart et al., 2019; Gama, 2023). Assessing soil carbon is critical for monitoring agricultural, livestock, and forestry activities, as well as restoration, reforestation, and soil ecosystem services in support of programs and activities aiming to promote and improve soil security.

The Brazilian National Forest Inventory (NFI) is a cornerstone program for monitoring the country's forest resources, providing data to support public policies and meet international climate commitments. A key component of the NFI is the systematic assessment of soil carbon, a primary indicator of soil health and a major factor in national carbon budgets. To support such programs, there is a growing demand for new procedures to assess soil carbon that are faster and more cost-effective than traditional techniques but still provide comparable accuracy (Viscarra Rossel et al., 2024). This is the case of visible-near infrared (Vis-NIR) spectroscopy, an analytical method that relies on the interaction of electromagnetic radiation in the Vis-NIR range (350–2500 nm) with the soil constituents, allowing to predict soil carbon content and other soil properties (Soriano-Disla et al., 2013; Viscarra Rossel et al., 2022). This method is non-destructive, rapid, inexpensive, and precise.

When infrared energy is emitted into the soil, the light spreads inside

* Corresponding author.
E-mail addresses: gustavo.vasques@embrapa.br (G.M. Vasques), leviluz2000@gmail.com (L.B. Luz), fabiano.balieiro@embrapa.br (F.C. Balieiro), monise.seas@gmail.com (M.A.F. Magalhães), telmoborges.florestal@gmail.com (T.B. Silveira Filho), marcelo.andrade@embrapa.br (M.T. Andrade).

https://doi.org/10.1016/j.soisec.2026.100236
Received 31 October 2025; Received in revised form 24 April 2026; Accepted 11 May 2026
Available online 14 May 2026

the sample, causing the bonds in the molecules that compose it to vibrate and stretch, absorbing part of this light, while some of the energy is diffusely reflected back. A Vis-NIR spectrometer measures the diffuse reflectance of light from the soil at narrow wavelength intervals across the Vis-NIR range, building a soil reflectance curve by plotting the soil reflectance against the wavelength. Since the soil constituents and other properties affect the reflectance behavior of the soil, this curve can be used to predict their amounts, including soil color (Viscarra and Lark, 2009), iron oxide contents (Viscarra Rossel et al., 2010), clay mineral content (Viscarra Rossel et al., 2011), and soil moisture (Baumann et al., 2022).

To use soil spectral curves to predict soil properties, one must first develop a soil spectral library (SSL), which is a collection of soil spectral curves accompanied by the values of the soil properties of interest measured using conventional laboratory analysis. Then, this SSL is used to calibrate a predictive model for a target soil property as a function of the spectral curves, for which different calibration methods can be used. Accordingly, different preprocessing transformations can be applied to the soil spectral curves to remove noise, improve signal-to-noise ratio, remove undesirable artifacts, smooth their variation, and transform the spectral data aiming to improve the quality of the soil property predictions (Dotto et al., 2017a, b). For instance, derivative transformations are frequently applied to spectral data using Savitzky-Golay polynomial filters (Savitzky and Golay, 1964).

As the most appropriate preprocessing transformations and multivariate calibration methods greatly depend on the soil characteristics, a common strategy to improve the quality of soil property predictions involves comparing different combinations of preprocessing transformations and calibration methods, and selecting those that yield the highest predictive performance (Rinnan et al., 2009; Rinnan, 2014). Thus, this study aims to build a soil carbon content prediction model from a statewide SSL for Rio de Janeiro state by comparing five spectral preprocessing transformations and five multivariate calibration methods. The model will support assessing soil carbon contents and stocks, as well as their changes across the Rio de Janeiro state between National Forest Inventory campaigns.

2. Materials and methods

2.1. Study area and sampling design

The study was conducted in the state of Rio de Janeiro, southeastern Brazil, with approximately 43,782 km^2. The state is characterized by humid tropical climate with high annual rainfall and average temperatures above 20 °C. The region has a complex topography and diverse vegetation cover, including remnants of the Atlantic Forest. The predominant soil types include Ferralsols (Oxisols), Acrisols (Ultisols), and Cambisols (Inceptisols), which are typically acidic and highly weathered, with variable organic matter content depending on the land use/land cover.

The soil samples used in the study were obtained from the first NFI of Rio de Janeiro state (SFB, 2018). The NFI employed a systematic sampling scheme with sampling sites distributed on a grid of about 20 × 20 km, with a total of 251 sampled sites. At each site, soil samples were collected using a borehole at depths of 0–20 and 30–50 cm, and analyzed for chemical and physical properties. A total of 373 samples (186 at 0–20 cm, and 187 at 30–50 cm) could be recovered from the stored NFI samples and were used in the study (Fig. 1).

As a reference, across the Rio de Janeiro state, soil carbon stocks vary from 9.1 to 96.7 Mg ha^{-1} at 0–20 cm, and from 7.0 to 63.8 Mg ha^{-1} at 30–50 cm (Vasques et al., 2025). Highest soil carbon stock values are found in the central, southwestern and northwestern mountainous regions of the state, and at the Paraíba do Sul River delta at the eastern boundary of the state. The same spatial trends are observed for the soil carbon contents measured at the NFI sites across the state (Fig. 1).

2.2. Laboratory soil analysis

The 376 soil samples were ground, sieved (2 mm), and dried at room temperature for 48 h. Soil carbon content was measured by dry combustion in a CHNS 2400 elemental analyzer (Perkin Elmer, Waltham, EUA). For measuring the Vis-NIR spectral reflectance, the samples were dried at 45 °C overnight for 15 h to harmonize the water content. Then, they were placed in a 10-cm-diameter Petri dish on an ASD Turntable (Malvern Panalytical, Malvern, United Kingdom) rotating at 22 rpm and

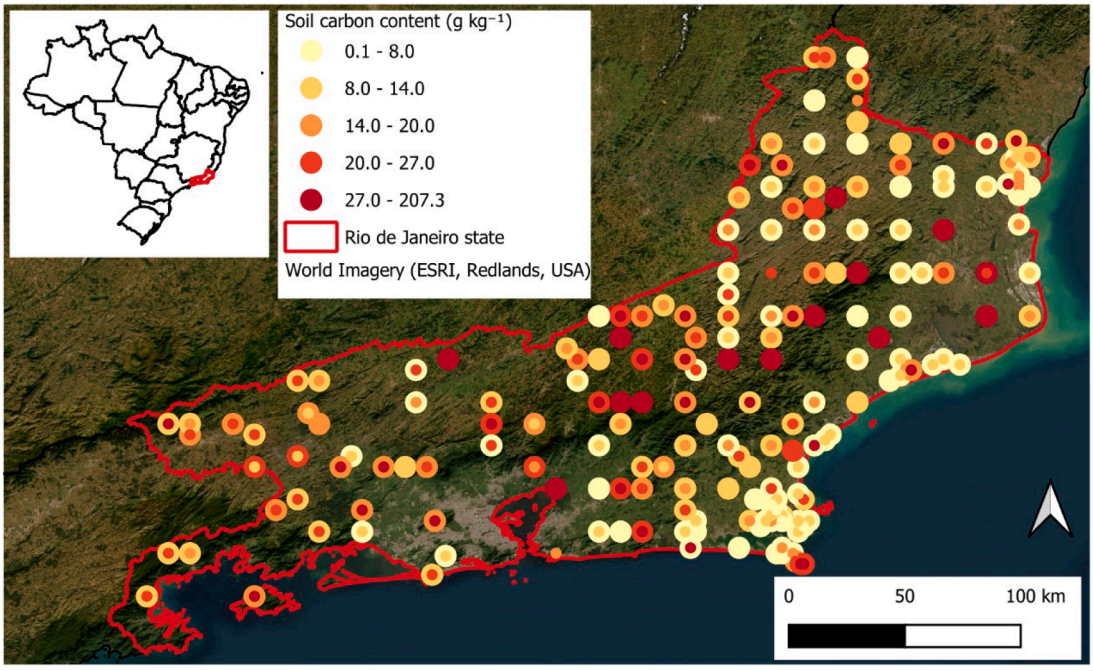

Fig. 1. Location of the study area – Rio de Janeiro state, Brazil – and soil carbon contents observed at the sampling sites. The smaller and larger dots refer to soil carbon contents at 0–20 and 30–50 cm, respectively.

illuminated by a 20 W halogen bulb. The soil Vis-NIR curves were acquired using an ASD FieldSpec 4 spectroradiometer (Malvern Panalytical, Malvern, United Kingdom), averaging 100 repetitions per sample. Spectralon® panel (Labsphere, North Sutton, USA) was used as a white reference (100 % reflectance), and scanned before each block of 10 readings.

2.3. Preprocessing of soil spectral curves

To ensure the development of a robust calibration model, a diagnostic analysis was performed to identify potential outliers that could exert undue leverage on the results. Using Oja distance (Oja, 1983; Zuo and Serfling, 2000), a multivariate distance metric suitable for high-dimensional data, 21 samples were identified and removed as spectrally distinct from the main population in the SSL, leaving 355 samples (174 at 0–20 cm, and 181 at 30–50 cm). The removal of these samples did not significantly compromise the soil spectral variability.

After outlier removal, the soil spectral curves were preprocessed using five spectral transformations (Table 1). First, to remove noise, the curves were smoothed across a moving window of 9 nm using the Savitzky-Golay algorithm with a third order polynomial (SGS; Savitzky and Golay, 1964). Then, four preprocessing transformations were applied to the smoothed spectral curves, including: Savitzky-Golay first derivative with a first-order polynomial and moving window of 9 nm (SGD); standard normal variate (SNV; Barnes et al., 1989); log (1/reflectance) (LOG); and continuum removal (CR; Clark and Roush, 1984). Spectral preprocessing was implemented using the prospectr package (Stevens and Ramirez-Lopez, 2025) in R (R Core Team, 2024).

2.4. Multivariate calibration methods

After the preprocessing transformations of soil spectral curves, the samples were split into calibration and validation sets by assigning 70 % of the sampling sites to calibration (248 samples), and the remaining 30 % to the validation set (107 samples), respectively. The calibration samples were used to train the soil carbon content prediction models from soil spectral curves, using five different multivariate calibration methods (Table 2), including: partial least squares regression (PLSR; Wold et al., 2001); elastic net regression (Tibshirani, 1996); cubist (Quinlan, 1993); random forest (RF; Breiman, 2001); and support vector machine (SVM; Cortes and Vapnik, 1995).

Partial least squares regression and elastic net are the two linear, parametric methods compared in the study. Partial least squares regression reduces data dimensionality by transforming the predictor variables into orthogonal (uncorrelated) latent variables, while preserving most of their information. It was implemented using the pls R package (Liland et al., 2024). Elastic net regression works by shrinking the coefficients of the predictor variables, combining ridge regression and least absolute shrinkage and selection operator (LASSO) regularizations by adjusting the alpha hyperparameter between 0 (ridge regression) and 1 (LASSO). The shrinkage strength is controlled by the lambda hyperparameter. Elastic net was implemented using the glmnet R package (Friedman et al., 2010).

Cubist and random forest are the two tree-based methods compared in the study. Cubist combines decision trees with linear regressions by, first, generating a set of rules to segment the data into smaller subsets (leaves), and then, fitting linear regression models to each of these subsets to make predictions. It was implemented in R using the Cubist package (Kuhn and Quinlan, 2024). Random forest builds a collection (a forest) of decision trees using a bootstrap aggregating (bagging) approach. For each individual tree, random selections of training and validation samples, as well as predictor variables, are used to build the tree and make predictions of the target variable. Then, the predictions from the individual trees are averaged to generate the final predictions. The randomForest R package (Liaw and Wiener, 2002) was used to derive the RF models.

The last method – SVM – works by separating samples using linear or non-linear boundaries (hyperplanes). In this study, linear hyperplanes were used. The cost (penalty) regularization hyperparameter (C) controls the trade-off between maximizing the margins of the hyperplanes that separate the samples leading to larger prediction errors and more generalization capacity (smaller C), and narrowing the margins forcing the model to minimize the prediction errors (larger C), which may lead to overfitting. It was implemented using the e1071 package (Meyer et al., 2023) in R. Together, these methods cover linear *versus* non-linear, and parametric *versus* non-parametric approaches, allowing for comprehensive performance comparison aiming at improving soil carbon content predictions.

For each calibration method, selected method hyperparameters (Table 2) were optimized to ensure the best prediction accuracy of soil carbon content. This was done with the caret package (Kuhn, 2008) in R via 10-fold cross-validation using the samples from the training set, with the best hyperparameters selected by minimizing the root mean square error (RMSE; Eq. 1) of cross-validation. Once the best hyperparameters were selected, a single model was derived from the training set using these hyperparameters, and then validated on the validation set, respectively. Model performance and prediction errors were evaluated using the RMSE and other error metrics, including: mean absolute error (MAE; Eq. 2); mean error (ME; Eq. 3); coefficient of determination (R^2; Eq. 4); and residual prediction deviation (RPD; Eq. 5). The smallest RMSE of cross-validation was used as criteria to select the best model for soil carbon content prediction, as it assesses the generalization capacity of the model, that is, its ability to make predictions on unknown samples.

Table 1
Preprocessing transformations of soil spectral curves.

Preprocessing transformation	Abbreviation
Savitzky-Golay smoothing with a third-order polynomial and moving window of 9 nm	SGS
Savitzky-Golay first derivative with a first-order polynomial and moving window of 9 nm	SGD
Standard normal variate	SNV
Log(1/reflectance)	LOG
Continuum removal	CR

Table 2
Multivariate calibration method optimized hyperparameters values.

Method	Hyperparameter	Values
Partial least squares regression	Number of latent variables (ncomp)	1,2,…,24
Elastic net	Mixing parameter (alpha)	0,0.1,0.2,…,1
	Regularization strength (lambda)	0.001,0.002, …,0.1
Cubist	Number of committees (committees)	1,2,…,10
	Number of nearest neighbors (neighbors)	1,2,…,9
Random forest	Number of trees (ntree)	100,200,…,500
	Number of split variables (mtry)	20,40,…,200
Support vector machine	Cost (C)	0.1,0.2,…,1

$$RMSE = \sqrt{\sum (y_i - \hat{y}_i)^2 \Big/ n} \quad \text{(Eq. 1)}$$

$$MAE = \frac{1}{n}\sum |y_i - \hat{y}_i| \quad \text{(Eq. 2)}$$

$$ME = \frac{1}{n}\sum (y_i - \hat{y}_i) \quad \text{(Eq. 3)}$$

$$R^2 = 1 - \sum (y_i - \hat{y}_i)^2 \Big/ \sum (y_i - \bar{y})^2 \quad \text{(Eq. 4)}$$

$$\mathrm{RPD} = SD_{val} \Big/ RMSE_v \sqrt{n/(n-1)} \quad \text{(Eq. 5)}$$

Where: $\widehat{y}_i$ = predicted values; y_i = observed values; $\overline{y}$ = mean of observed values; n = sample size; SD_{val} = standard deviation of the observed values in the validation set; $RMSE_v$ = root mean square error of validation.

To further improve soil carbon content predictions, an ensemble model was derived by fitting a generalized linear model with a Gaussian error distribution and identity link function, using as input variables the predictions from the best model obtained by each method, respectively. This model ensemble approach aimed to produce more robust soil carbon predictions by harnessing the advantages of each method while minimizing their individual limitations. Training and validation errors were calculated for the ensemble model, and compared against the individual models.

2.5. Variable importance

In PLSR, the variable importance was calculated as its proportional contribution to the reduction in sums of squares. It was calculated as the weighted sum of the absolute value of the regression coefficient, with the weights derived from the reduction in sums of squares over the PLSR components. In elastic net, since the variable coefficients are shrunk towards zero, effectively performing automatic variable selection, variable importance was directly inferred from the absolute value of the coefficients, with larger coefficients indicating more important variables. In cubist, variable importance was calculated by how frequently each variable is included in decision rules along the trees and in the linear regression models at the leaves.

In RF, variable importance was calculated as follows. First, the prediction accuracy on the out-of-bag samples is calculated. Then, this process is repeated after randomly permuting the values of each predictor variable. Variable importance is assessed by the difference in prediction accuracy between these two cases averaged across all trees and normalized by the standard error. Finally, in SVM, variable importance was calculated as the absolute values of the coefficients.

3. Results and discussion

3.1. Descriptive statistics

Soil carbon contents varied from 0.1 to 49.5 g kg^{-1}, with a median of 13.2 g kg^{-1}, mean of 15.5 g kg^{-1} and standard deviation of 9.9 g kg^{-1}, showing a slightly right-skewed distribution (Table 3). Soil carbon content variation in the training set was similar to, and encompassed the variation of the validation set. This large variation in soil carbon content directly reflects the heterogeneity of the soils in the state of Rio de Janeiro, ranging from carbon-poor sandy soils to carbon-rich organic soils.

The soil spectral curves show distinct spectral signatures among the samples, characterized by variations in albedo, shape, and absorption features (Fig. 2a, b). Despite these differences, the soils exhibited notable similarities in the general shape of their spectral curves and in the position of key absorption bands, mainly associated with dominant components including minerals, organic matter, and moisture. Preprocessing transformations considerably changed the shape of the original spectral curves (Fig. 2c, d, e, f), marking and strengthening distinct absorption features that were captured by the different calibration methods.

3.2. Performance of individual prediction models

Based on the RMSE of cross-validation ($RMSE_{cv}$), the cubist model fit to LOG data derived the best soil carbon content predictions, with a $RMSE_{cv}$ of 5.1 g kg^{-1}, followed by PLSR-LOG and SVM-SNV models, both with a $RMSE_{cv}$ of 5.4 g kg^{-1} (Table 4). The cross-validation approach is particularly valuable for addressing the heterogeneity of the statewide soil samples, providing more conservative and reproducible error estimates than a single external validation set (Kuhn and Johnson, 2013).

Based on the average $RMSE_{cv}$, the performance of calibration methods was as follows, from best to worst ($RMSE_{cv}$ in parentheses, in g kg^{-1}): SVM (5.7) > cubist (5.8) > PLSR (5.9) > elastic net (6.1) > RF (6.5). Accordingly, the performance of preprocessing transformations was as follows: LOG (5.8) > SNV (5.9) > SGS (6.0) = SGS (6.0) > CR (6.4). Comparatively, SVM, cubist and PLSR had similar performance, whereas RF degraded the soil carbon content predictions, with RF-LOG and RF-SGS showing the poorest performance among all method and preprocessing transformation combinations. Among preprocessing transformations, CR showed poorer performance, especially in PLSR, elastic net, and SVM, where it derived the worst predictions. One possible explanation is that CR transformation suppressed chemically meaningful variance removing important features, patterns and correlations from the spectral curves that could otherwise be exploited by the multivariate methods.

Overall, PLSR demonstrated consistent prediction accuracy across preprocessing transformations, except CR, where its performance notably declined. Based on the $RMSE_{cv}$, the best PLSR model was obtained using LOG transformation, similarly to cubist. This model achieved a RMSE of external validation ($RMSE_v$) of 5.8 g kg^{-1}, and a RPD of 1.72 (Table 4), meaning that the standard deviation of the errors is considerably smaller than the standard deviation of the observed soil carbon content, indicating that the model can be used for soil carbon content prediction with moderate accuracy (Chang et al., 2001). This is confirmed by the predicted *versus* observed plot (Fig. 3), which shows overall consistent predictions of soil carbon content along the 1:1 correlation line, with a few exceptions.

Overall, elastic net was superior to RF, but slightly inferior to all other calibration methods. The smallest $RMSE_{cv}$ (5.6 g kg^{-1}) was obtained from SNV data. The elastic net results align with the results from a recent study in Pernambuco, northeastern Brazil, where elastic net (LASSO, specifically) achieved a $RMSE_{cv}$ of 4.1 g kg^{-1} predicting soil organic carbon content (Santos et al., 2023). The best elastic net external validation results were derived from SGS data, and correspond to a $RMSE_v$ of 6.4 g kg^{-1}, and RPD of 1.55. Previous Vis-NIR spectroscopy studies have also shown that preprocessing transformation of soil spectral curves enhances model accuracy (Dotto et al., 2017a; Moura-Bueno et al., 2018). The predicted *versus* observed plot (Fig. 4) shows a trend of underestimating large values and overestimating small ones, which is common for regression models, especially linear-based ones.

Cubist obtained the best overall model among all method and preprocessing transformation combinations. This best model, which used LOG data as input, also performed best in external validation, obtaining a $RMSE_v$ of 5.8 g kg^{-1}, and RPD of 1.72. This is the largest RPD among all models, which was also achieved by the PLSR-LOG model. This shows

Table 3
Descriptive statistics of soil carbon content (g kg^{-1}).

Sample set	Sample size	Mean	SD	Median	Minimum	Maximum	Skewness	Kurtosis
Whole	355	15.5	9.9	13.2	0.1	49.5	1.09	1.15
Training	248	15.4	9.8	13.3	0.1	49.5	1.16	1.47
Validation	107	15.9	10.1	12.8	2.7	49.1	0.93	0.43

SD = standard deviation.

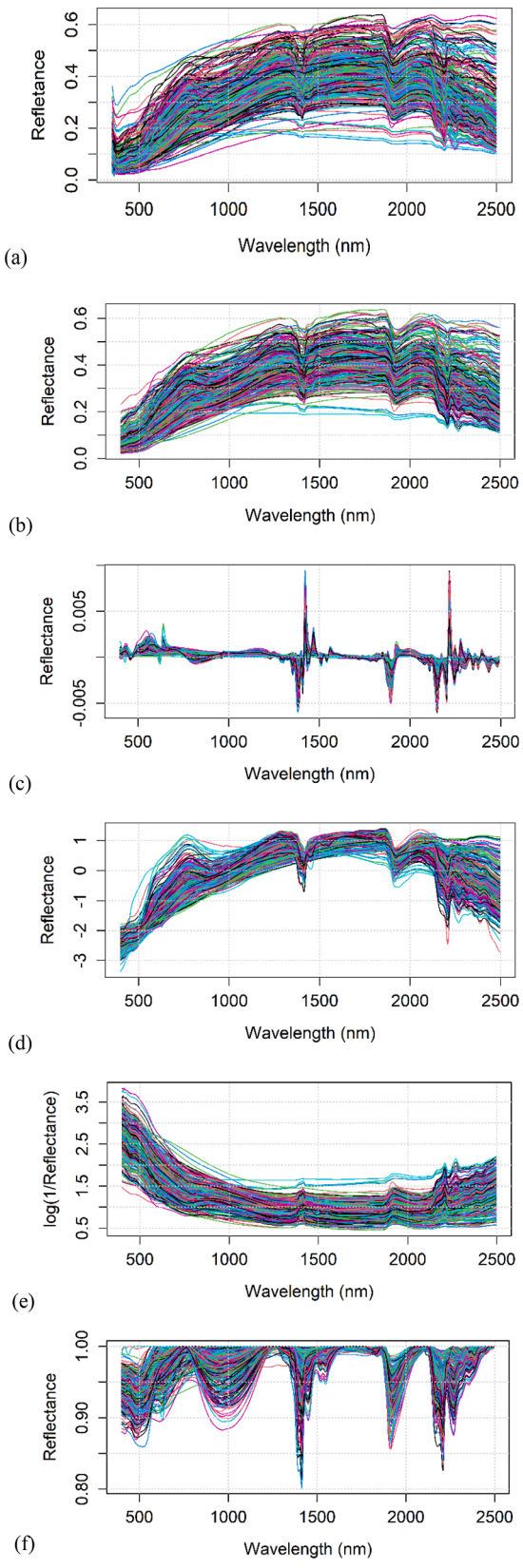

Fig. 2. Preprocessed spectral curves: (a) Original curves; (b) Savitzky-Golay smoothing with a third-order polynomial and moving window of 9 nm; (c) Savitzky-Golay first derivative with a first-order polynomial and moving window of 9 nm; (d) Standard normal variate; (e) Log(1/reflectance); and (f) Continuum removal.

the good generalization capacity of the cubist-LOG model to make soil carbon content predictions across the Rio de Janeiro state, as also observed in Fig. 5. A key strength of cubist lies in its consistent use of three or five committees, which refine and stabilize predictions by boosting rule-based ensembles, reducing overfitting. On the other hand, larger datasets might require more committees to capture complex patterns, which may compromise the model's efficiency in terms of processing time, and its ability to generalize predictions.

Since minimizing the RMSE of cross-validation was used to both optimize the model hyperparameters and select the best prediction method, this may have led to model overfitting in the training phase, which possibly degraded the predictions in external validation. In effect, the cubist models had very large R^2 of training, suggesting overfitting, which was also the case for RF models. On the other hand, their error metrics from cross-validation and external validation are larger, and comparable to those from the other methods.

These results corroborate the findings of Dematte et al. (2019) who evaluated the potential of a Vis-NIR spectral library for predicting soil organic carbon content, clay content, and other soil properties across Brazil using cubist. They found a $RMSE_v$ of 6.9 g kg^{-1} using a nationwide dataset. A similar study by Moura-Bueno et al. (2021) also using cubist achieved RMSE of 5.3 to 6.6 g kg^{-1} for soil organic carbon content prediction in Brazilian subtropical soils, using soil Vis-NIR data or combining Vis-NIR, spectral classes and environmental data. In another study in soils from the Brazilian Cerrado to Atlantic Forest transition, cubist obtained the smallest $RMSE_v$ (4.9 g kg^{-1}) compared to PLSR and RF, agreeing to this study. However, across preprocessing transformations, cubist showed a similar performance to PLSR ($RMSE_v$ of 5.6 to 7.0 g kg^{-1}) and RF ($RMSE_v$ of 5.5 to 6.8 g kg^{-1}) (Mendes et al., 2021).

Random forest models delivered the poorest soil carbon content predictions among the multivariate calibration methods, with both cross-validation and external validation metrics consistently showing poor model performance. The best-performing RF model used SGD and obtained a $RMSE_{cv}$ of 6.1 g kg^{-1}, comparable to other methods. However, all other preprocessing transformations lagged behind the alternative methods, with $RMSE_{cv}$ consistently larger than 6.4 g kg^{-1}. In external validation, the best RF model also used SGD data as input, reaching $RMSE_v$ (6.9 g kg^{-1}) and RPD (1.44) metrics worse than all other methods, except SVM. Along the same lines, the point cloud is more dispersed in the predicted *versus* observed plot (Fig. 6) compared to the other methods, except SVM, whose external validation was the poorest among the calibration methods.

The RF models consistently underperformed compared to the PLSR, cubist, and SVM models. This was also the case in Mendes et al. (2021), where RF was outperformed by cubist predicting soil pH, and carbon, clay and sand contents. Although RF effectively handles regression tasks, its structure can sometimes lead to decreased model generalization capacity, especially when applied to highly heterogeneous datasets. The lack of parametric and linear structures in RF, compared to other methods such as PLSR, elastic net, or cubist, may hinder its ability capture strong correlations among soil carbon and soil spectral reflectance. Moreover, RF model outputs are less interpretable than linear models (Breiman, 2001).

Support vector machine models exhibited solid performance across preprocessing transformations in cross-validation, with the best $RMSE_{cv}$ among all the methods. However, in external validation, it failed to provide strong predictions, achieving the poorest average $RMSE_v$ (7.9 g kg^{-1}) and average RPD (1.27) among the methods. This is clear from the widespread and oval-shaped distribution of predicted *versus* observed validation samples (Fig. 7). In comparison, better results were obtained by Terra et al. (2015), who predicted soil organic carbon content using SVM from 1259 samples from four Brazilian states. They obtained a $RMSE_v$ of 0.16 log10(g kg^{-1}) compared to the 7.3 g kg^{-1} obtained from the SVM-SGS model, the best SVM model in external validation in this study. Support vector machine showed comparable performance to PLSR for predicting soil clay, and organic matter contents in 7172

Table 4
Optimized method hyperparameters and prediction error metrics.

Method	Preprocessing	Optimal hyperparameters	$RMSE_{cv}$	MAE_{cv}	R^2_{cv}	$RMSE_t$	MAE_t	ME_t	R^2_t	$RMSE_v$	MAE_v	ME_v	R^2_v	RPD_v
PLSR	SGS	ncomp=16	5.7	4.3	0.66	5.1	3.7	0	0.72	5.9	4.5	−0.0	0.65	1.68
PLSR	SGD	ncomp=11	5.8	4.3	0.67	5.0	3.7	0	0.73	6.1	4.7	−0.2	0.63	1.64
PLSR	SNV	ncomp=16	5.8	4.3	0.65	5.2	3.7	0	0.71	9.7	5.1	−1.1	0.37	1.03
PLSR	LOG	ncomp=16	5.4	4.0	0.71	4.7	3.4	0	0.76	5.8	4.1	0.2	0.66	1.72
PLSR	CR	ncomp=27	6.7	5.0	0.57	4.4	3.3	0	0.79	9.6	6.2	−0.7	0.30	1.05
Elastic net	SGS	alpha=1, lambda=0.05	6.1	4.5	0.61	5.9	4.2	0	0.63	6.4	4.8	−0.1	0.60	1.55
Elastic net	SGD	alpha=0.1, lambda=0.1	5.6	4.3	0.67	4.3	3.2	0	0.80	6.5	4.8	−0.2	0.57	1.53
Elastic net	SNV	alpha=0.9, lambda=0.05	6.0	4.4	0.63	5.8	4.1	0	0.64	7.0	5.0	−0.4	0.51	1.42
Elastic net	LOG	alpha=1, lambda=0.04	6.1	4.4	0.59	5.8	4.1	0	0.65	6.5	4.7	0.2	0.58	1.54
Elastic net	CR	alpha=1, lambda=0.03	6.7	5.1	0.53	5.4	4.0	0	0.69	7.2	5.6	−0.5	0.49	1.39
Cubist	SGS	committees=3, neighbors=5	5.7	4.0	0.66	0.4	0.1	0	1.00	7.2	5.4	−1.3	0.54	1.40
Cubist	SGD	committees=5, neighbors=5	6.4	4.8	0.59	0.3	0.1	0	1.00	7.5	5.8	−1.8	0.48	1.34
Cubist	SNV	committees=5, neighbors=5	5.7	4.1	0.65	0.2	0.1	0	1.00	6.6	4.8	−0.9	0.58	1.52
Cubist	LOG	committees=5, neighbors=0	5.1	3.5	0.73	3.4	2.4	0.1	0.89	5.8	4.1	−0.1	0.66	1.72
Cubist	CR	committees=5, neighbors=5	6.0	4.4	0.58	0.4	0.1	0	1.00	7.3	5.4	−0.8	0.48	1.38
RF	SGS	ntree=250, mtry=50	6.7	5.0	0.53	2.7	1.9	0	0.95	7.8	6.1	0.1	0.39	1.28
RF	SGD	ntree=750, mtry=50	6.1	4.6	0.65	2.4	1.7	−0.1	0.97	6.9	5.4	−0.7	0.55	1.44
RF	SNV	ntree=750, mtry=50	6.4	4.9	0.56	2.6	1.9	−0.1	0.95	7.6	6.0	−0.1	0.43	1.32
RF	LOG	ntree=750, mtry=50	6.8	5.0	0.52	2.7	1.9	−0.1	0.94	7.8	6.1	−0.1	0.40	1.29
RF	CR	ntree=750, mtry=50	6.5	4.9	0.57	2.6	1.9	−0.1	0.96	7.5	6.2	−0.7	0.47	1.34
SVM	SGS	$C = 1$	5.7	4.1	0.67	6.1	4.0	0.5	0.63	7.3	5.5	0.7	0.47	1.36
SVM	SGD	$C = 0.1$	5.9	4.3	0.64	7.1	4.7	1.4	0.63	8.2	6.3	1.7	0.45	1.22
SVM	SNV	$C = 0.9$	5.4	3.9	0.71	5.6	3.6	1.0	0.69	7.6	5.9	0.8	0.43	1.32
SVM	LOG	$C = 1$	5.5	4.0	0.69	5.8	3.9	0.7	0.68	7.6	5.6	0.9	0.43	1.31
SVM	CR	$C = 0.2$	5.9	4.4	0.64	7.0	4.6	1.2	0.57	8.6	6.7	1.1	0.30	1.16

RMSE = Root mean square error; MAE = Mean absolute error; R^2 = Coefficient of determination; ME = Mean error; RPD = Residual prediction deviation; cv = cross-validation; t = training; v = validation; PLSR = Partial least squares regression; RF = Random forest; SVM = Support vector machine; SGS = Savitzky-Golay smoothing; SGD = Savitzky-Golay first derivative; SNV = Standard normal variate; LOG = Log(1/reflectance); CR = Continuum removal.

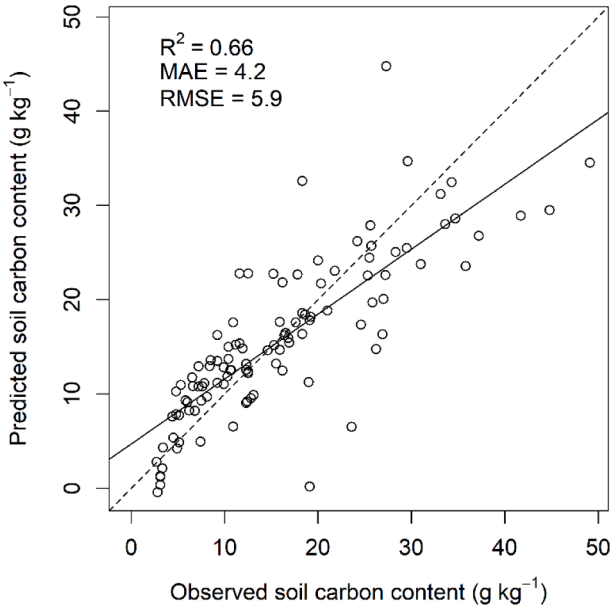

Fig. 3. Predicted *versus* observed soil carbon content of external validation for the partial least squares regression–log(1/reflectance) model. R^2 = Coefficient of determination; MAE = Mean absolute error; RMSE = Root mean square error.

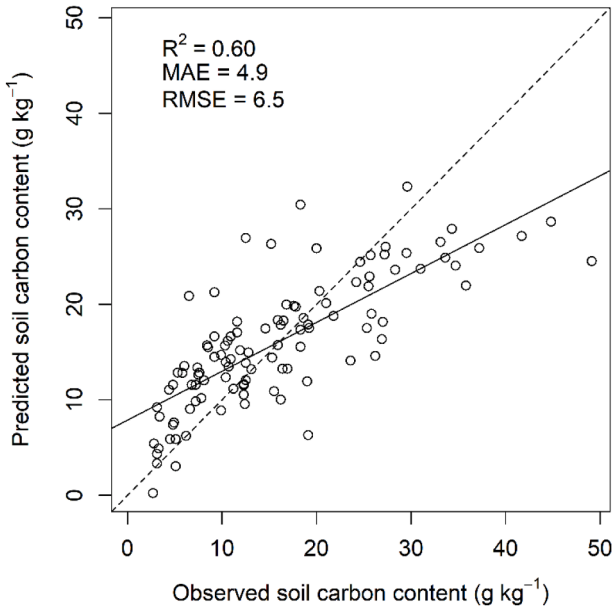

Fig. 4. Predicted *versus* observed soil carbon content of external validation for the elastic net–Savitzky-Golay smoothing model. R^2 = Coefficient of determination; MAE = Mean absolute error; RMSE = Root mean square error.

samples from different regions of Brazil (Araújo et al., 2014).

3.3. Performance of the ensemble model

The ensemble model training R^2, MAE, and RMSE were 0.98, 1.1 g kg^{-1}, and 1.5 g kg^{-1}, respectively. In external validation, the ensemble model produced slightly worse predictions than the PLSR, elastic net, and cubist best models, but outperformed RF and SVM, with R^2 of 0.52, RPD of 1.43, MAE of 5.3 g kg^{-1}, and RMSE of 7.0 g kg^{-1}. The predicted *versus* observed plot of external validation (Fig. 8) shows a tendency of underestimation of large soil carbon values, and overestimation of small ones, similar to the individual models. Moreover, like the other models, the variance of prediction errors is not constant across the soil carbon range. In this sense, the ensemble model did not fix the individual model drawbacks when predicting soil carbon on unknown samples.

Compared to individual models, the model ensemble approach takes one more step in data processing, making it less parsimonious, and prone to error propagation. Thus, although the training errors were very good, probably due to overfitting, the large validation errors combined to the

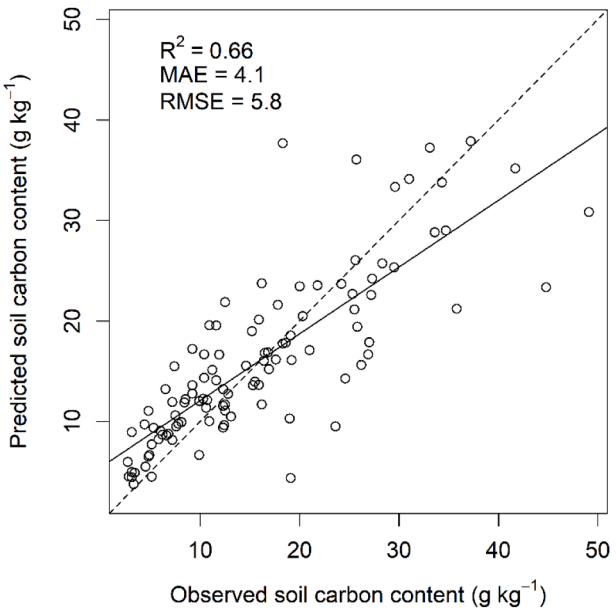

Fig. 5. Predicted *versus* observed soil carbon content of external validation for the cubist–log(1/reflectance) model. R^2 = Coefficient of determination; MAE = Mean absolute error; RMSE = Root mean square error.

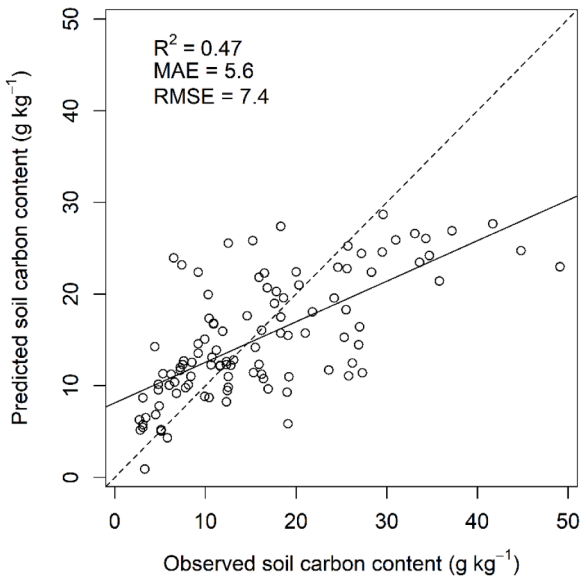

Fig. 7. Predicted *versus* observed soil carbon content of external validation for the support vector machine–Savitzky-Golay smoothing model. R^2 = Coefficient of determination; MAE = Mean absolute error; RMSE = Root mean square error.

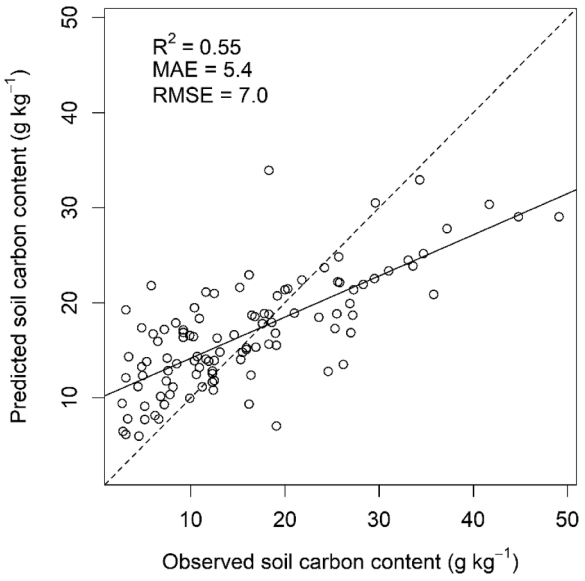

Fig. 6. Predicted *versus* observed soil carbon content of external validation for the random forest–Savitzky-Golay first derivative model. R^2 = Coefficient of determination; MAE = Mean absolute error; RMSE = Root mean square error.

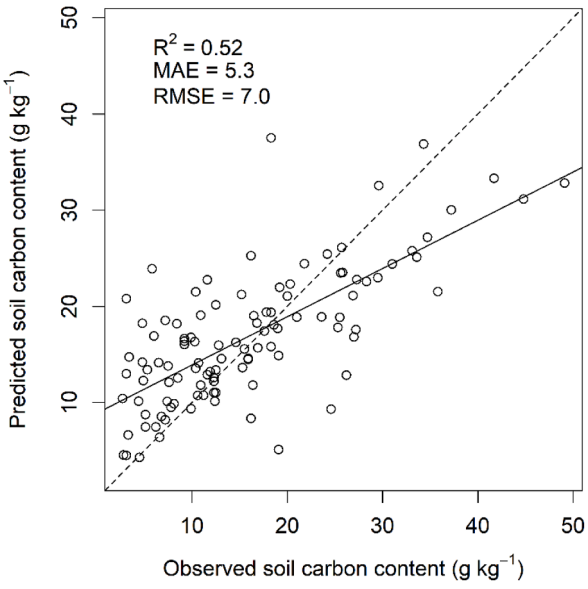

Fig. 8. Predicted *versus* observed soil carbon content of external validation for the ensemble model. R^2 = Coefficient of determination; MAE = Mean absolute error; RMSE = Root mean square error.

abovementioned limitations make the ensemble model less attractive for a practical routine use.

3.4. Variable importance

The performance of the soil carbon content prediction models is primarily driven by specific wavelengths identified as important variables, which include the spectral bands most relevant for soil carbon prediction. The most important wavelengths to predict soil carbon content varied by calibration method, appearing in different regions of the Vis-NIR range (Fig. 9). In the PLSR model, the most important variables were observed near 1400 and 2500 nm, whereas they were concentrated in the 1400 nm region in the SVM model. The band around 1400 nm is typically associated with O–H stretching related to residual water retained in the soil, while the range from 2200 to 2500 nm is linked to molecular vibrations including O–H stretching and metal–OH bending particularly associated with Al–OH and clay minerals such as kaolinite and smectites (Viscarra Rossel et al., 2005; Stenberg et al., 2010).

For the elastic net model, the most important wavelengths for soil carbon content prediction were located around 460, between 1400 and 1650, and at 2500 nm. The reflectance at 460 nm is likely related to goethite, an iron oxide common in tropical soils. The 1400–1650 nm region contains absorption features related to O–H functional groups (Stenberg et al., 2010; Terra et al., 2015), while 2500 nm may be related to the last bands of the soil spectral curves, which showed greater

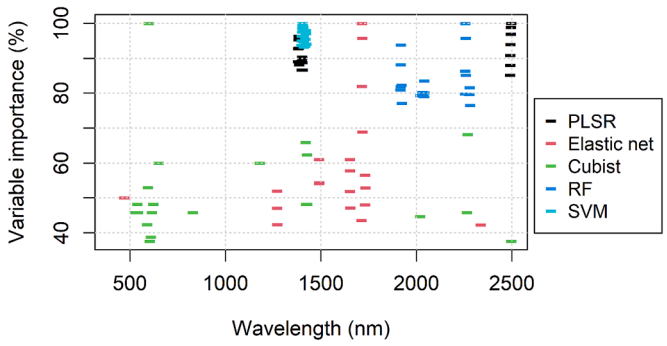

Fig. 9. Twenty most important variables in the best soil carbon content prediction model for each calibration method, respectively. PLSR = Partial least squares regression; RF = Random forest; SVM = Support vector machine.

spectral variation among the soil samples. For the cubist model, the variables with the highest importance clustered in the 1400 and 2200 nm regions, and in the visible region between 500 and 650 nm, the latter influenced by the presence of different minerals and organic matter with different colors, although specific assignments in this region are less commonly emphasized in Vis-NIR soil spectroscopy because they have limited chemistry specificity compared to bands with known absorption features from molecular groups (Stenberg et al., 2010).

The RF model highlighted variables around 1850, 1950, and 2200 nm. The 1850 nm band is often attributed to carbonate (–CO_3) overtones, while 1950 nm is strongly associated with water absorption. The 2200 nm region is related to C–H and C–O stretch combinations, as well as N–H stretch interactions (Burns and Ciurczak, 2007). In addition, multiple models consistently identified important bands at 1100, and 1900 nm, which are the absorption regions of functional molecular groups including carboxyl (1100 nm), methyl (1100 nm), and hydroxyl (1900 nm) (Stenberg et al., 2010; Burns and Ciurczak, 2007).

Similar Vis-NIR bands were selected in the same spectral regions by two variable selection methods to predict soil organic carbon and particle size fractions in southern Brazil using PLSR and SVM (Dotto et al., 2017b). In Florida, USA, important bands for soil total carbon and soil organic carbon prediction by PLSR were observed at around 1000, 1850 and 1920 nm (Knox et al., 2015), and 500–700, 1400, and 1800–2400 nm (Vasques et al., 2010), coinciding with some of the important variables observed in this study. In addition, using a global, and local Vis-NIR datasets to predict soil organic carbon in different parts of the world, Viscarra Rossel et al. (2024) identified important bands and spectral regions similar to those observed in this study. For the Brazilian site, they concentrated around 600, 800, 1400–1500, 1900, and 2000–2200 nm.

4. Conclusions

Soil carbon plays a crucial role as a key indicator of soil functionality, and is essential for monitoring and promoting soil health and long-term soil security to benefit the society. Visible-near infrared spectroscopy shows potential for soil carbon content prediction, with most multivariate methods yielding reasonable models, with RPD > 1.5. Among the methods evaluated, on average, SVM produced the models with the smallest $RMSE_{cv}$. However, the SVM models failed in external validation, with limited capacity to make soil carbon content predictions on unknown samples. Cubist, followed by elastic net, had the best compromise between $RMSE_{cv}$ and $RMSE_{v}$, showing robust prediction performance in cross-validation as well as in external validation. As such, they can be used for carbon content prediction in the soils from the state of Rio de Janeiro, Brazil, within landscapes and geological contexts similar to those studied.

Preprocessing soil spectral curves is recommended for improving soil carbon content predictions, with log(1/reflectance) transformation emerging as a strong approach to prepare soil spectral curves before multivariate chemometric calibration. The combination of log(1/reflectance) preprocessing followed by cubist calibration produced the best soil carbon content predictions, with the smallest cross-validation, and external validation errors, being well suited for use in future soil carbon assessment projects in Rio de Janeiro. On the other hand, model ensemble is not recommended for assessing soil carbon in future projects as it did not improve predictions while adding an extra step in data processing.

This study has direct implications for global soil security and climate change challenges, particularly within the context of the United Nations Framework Convention on Climate Change (UNFCCC), and United Nations Convention to Combat Desertification (UNCCD). These conventions require countries to monitor, report, and verify changes in soil carbon stocks, a core indicator of land health and security. However, the feasibility of such programs is often constrained by the high costs and slow turnaround of conventional laboratory analyses, which can be overcome by adopting faster and reliable soil carbon assessment methods like the one presented in this study.

CRediT authorship contribution statement

Gustavo M. Vasques: Writing – review & editing, Writing – original draft, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Levi B. Luz:** Writing – original draft, Formal analysis, Data curation. **Fabiano C. Balieiro:** Supervision, Resources, Data curation, Conceptualization. **Monise A. F. Magalhães:** Resources, Funding acquisition. **Telmo B. Silveira Filho:** Resources, Project administration, Funding acquisition. **Marcelo T. Andrade:** Writing – review & editing, Data curation.

Declaration of competing interest

The authors declare no conflicts of interest.

Acknowledgements

To the Serviço Florestal Brasileiro for providing the soil samples used in the study. To Tatiane M. Araújo, Bárbara C. Andrade, and Lygia C. S. Roque for their help preparing and analyzing the soil samples. And to the two anonymous reviewers that helped to considerably improve this paper. Institutional support was provided by Secretaria de Estado do Ambiente e Sustentabilidade do Rio de Janeiro and Embrapa Cooperation Agreement 25100.23/0109–8. Funding was provided by the Climate Group Under2 Coalition Future Fund, Mata Atlântica Fund and Embrapa (grant number 20.24.00.027.00.00).

Data availability

The authors do not have permission to share data.

References

Araújo, S.R., Wetterlind, J., Demattê, J.A.M., Stenberg, B., 2014. 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 65, 718–729. https://doi.org/10.1111/ejss.12165.

Banwart, S.A., Nikolaidis, N.P., Zhu, Y.-G., Peacock, C.L., Sparks, D.L., 2019. Soil functions: connecting Earth's critical zone. Annu. Rev. Earth. Planet. Sci 47, 333–359. https://doi.org/10.1146/annurev-earth-063016-020544.

Barnes, R.J., Dhanoa, M.S., Lister, S.J., 1989. Standard normal variate transformation and de-trending of near infrared diffuse reflectance spectra. Appl. Spectrosc 43, 772–777. https://doi.org/10.1366/0003702894202201.

Baumann, M., Gasparri, I., Buchadas, A., Oeser, J., Meyfroidt, P., Levers, C., Romero-Muñoz, A., Waroux, Y.P., Müller, D., Kuemmerle, T., 2022. Frontier metrics for a process-based understanding of deforestation dynamics. Environ. Res. Lett. 17, 095010. https://doi.org/10.1088/1748-9326/ac8b9a.

Breiman, L., 2001. Random forests. Mach. Learn 45, 5–32. https://doi.org/10.1023/a:1010933404324.

Burns, D.A., Ciurczak, E.W. (Eds.), 2007. Handbook of Near-Infrared Analysis. CRC Press, Boca Raton, USA. https://doi.org/10.1201/9781420007374.
Chang, C., Laird, D.A., Mausbach, M.J., Hurburgh, C.R., 2001. Near-infrared reflectance spectroscopy–principal components regression analyses of soil properties. Soil. Sci. Soc. Am. J. 65, 480–490. https://doi.org/10.2136/sssaj2001.652480x.
Clark, R.N., Roush, T.L., 1984. Reflectance spectroscopy: quantitative analysis techniques for remote sensing applications. J. Geophys. Res.:. Solid. Earth 89, 6329–6340. https://doi.org/10.1029/jb089ib07p06329.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn 20, 273–297. https://doi.org/10.1007/BF00994018.
Demattê, J.A.M., Dotto, A.C., Paiva, A.F.S., Sato, M.V., Dalmolin, R.S.D., Araújo, M.S.B., Silva, E.B., Nanni, M.R., Caten, A., Noronha, N.C., Lacerda, M.P.C., Araújo Filho, J. C., Rizzo, R., Bellinaso, H., Francelino, M.R., Schaefer, C.E.G.R., Vicente, L.E., Santos, U.J., Sampaio, E.V.S.B., Menezes, R.S.C., Souza, J.J.L.L., Abrahão, W.A.P., Coelho, R.M., Grego, C.R., Lani, J.L., Fernandes, A.R., Gonçalves, D.A.M., Silva, S.H. G., Menezes, M.D., Curi, N., Couto, E.G., Anjos, L.H.C., Ceddia, M.B., Pinheiro, E.F. M., Grunwald, S., Vasques, G.M., Marques Júnior, J., Silva, A.J., Barreto, M.C.V., Nóbrega, G.N., Silva, M.Z., Souza, S.F., Valladares, G.S., Viana, J.H.M., Terra, F.S., Horák-Terra, I., Fiorio, P.R., Silva, R.C., Frade Júnior, E.F., Lima, R.H.C., Alba, J.M. F., Souza Junior, V.S., Brefin, M.L.M.S., Ruivo, M.L.P., Ferreira, T.O., Brait, M.A., Caetano, N.R., Bringhenti, I., Sousa Mendes, W., Safanelli, J.L., Guimarães, C.C.B., Poppiel, R.R., Souza, A.B., Quesada, C.A., Couto, H.T.Z., 2019. The Brazilian Soil Spectral Library (BSSL): a general view, application and challenges. Geoderma 354, 113793. https://doi.org/10.1016/j.geoderma.2019.05.043.
Dotto, A.C., Dalmolin, R.S.D., Caten, A., Grunwald, S., 2018. 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 314, 262–274. https://doi.org/10.1016/j.geoderma.2017.11.006.
Dotto, A.C., Dalmolin, R.S.D., Grunwald, S., Caten, A., Pereira Filho, W., 2017b. Two preprocessing techniques to reduce model covariables in soil property predictions by vis-NIR spectroscopy. Soil. Tillage. Res. 172, 59–68. https://doi.org/10.1016/j.still.2017.05.008.
Friedman, J., Tibshirani, R., Hastie, T., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw 33, 1–22. https://doi.org/10.18637/jss.v033.i01.
Gama, J.T., 2023. The role of soils in sustainability, climate change, and ecosystem services: challenges and opportunities. Ecologies 4, 552–567. https://doi.org/10.3390/ecologies4030036.
Knox, N.M., Grunwald, S., McDowell, M.L., Bruland, G.L., Myers, D.B., Harris, W.G., 2015. Modelling soil carbon fractions with visible near-infrared (VNIR) and mid-infrared (MIR) spectroscopy. Geoderma 239-240, 229–239. https://doi.org/10.1016/j.geoderma.2014.10.019, 229–239.
Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer, New York, USA. https://link.springer.com/book/10.1007/978-1-4614-6849-3.
Kuhn, M., Quinlan, R., 2024. Cubist: Rule- and instance-Based Regression Modeling. R package version 0.4.4. https://CRAN.R-project.org/package=Cubist.
Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw 28, 1–26. https://doi.org/10.18637/jss.v028.i05.
Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R. News 2, 18–22.
Liland, K., Mevik, B., Wehrens, R.. pls: Partial least squares and principal component regression. R package version 2.8-5. https://CRAN.R-project.org/package=pls.
Mendes, W.S., Boechat, C.L., Gualberto, A.V.S., Barbosa, R.S., Silva, Y.J.A.B., Saraiva, P. C., Sena, A.F.S., Duarte, L.S.L., 2021. Soil spectral library of Piauí state using machine learning for laboratory analysis in northeastern Brazil. Rev. Bras. Ciênc. Solo 45 (e0200115). https://doi.org/10.36783/18069657rbcs20200115.
Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., 2023. e1071: misc functions of the Department of Statistics, Probability Theory Group (formerly: E1071), TU Wien. R package version 1.7-14. https://CRAN.R-project.org/package=e1071.
Moura-Bueno, J.M., Dalmolin, R.S.D., Caten, A., Dotto, A.C., Demattê, J.A.M., 2019. Stratification of a local VIS-NIR-SWIR spectral library by homogeneity criteria yields more accurate soil organic carbon predictions. Geoderma 337, 565–581. https://doi.org/10.1016/j.geoderma.2018.10.015.
Moura-Bueno, J.M., Simão, R., Horst-Heinen, Taciara Zborowski, Grunwald, S., Caten, A. ten, 2021. Environmental covariates improve the spectral predictions of organic carbon in subtropical soils in southern Brazil. Geoderma 393, 114981. https://doi.org/10.1016/j.geoderma.2021.114981. –114981.
Oja, H., 1983. Descriptive statistics for multivariate distributions. Stat. Probab. Lett 1, 327–332. https://doi.org/10.1016/0167-7152(83)90054-8.
Quinlan, J.R., 1993. Combining instance-based and model-based learning. In: Proceedings of the Tenth International Conference on Machine Learning, pp. 236–243. https://doi.org/10.5555/3091529.3091560.
R Core Team, 2024. R: A language and Environment For Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org.
Rinnan, A., Berg, F., Engelsen, S.B., 2009. Review of the most common pre-processing techniques for near infrared spectra. Trends. Anal. Chem. 28, 1201–1222. https://doi.org/10.1016/j.trac.2009.07.007.
Rinnan, A., 2014. Pre-processing in vibrational spectroscopy – When, why and how. Anal. Methods 6, 7124–7129. https://doi.org/10.1039/c3ay42270d.
Santos, E.P., Moreira, M.C., Fernandes-Filho, E.I., Demattê, J.A.M., Santos, U.J., Silva, D. D., Cruz, R.R.P., Moura-Bueno, J.M., Santos, I.C., Sampaio, E.V.S.B., 2023. Improving the generalization error and transparency of regression models to estimate soil organic carbon using soil reflectance data. Ecol. Inf. 77, 102240. https://doi.org/10.1016/j.ecoinf.2023.102240.
Savitzky, A., Golay, E., 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem 36, 1627–1639. https://doi.org/10.1021/ac60214a047.
SFB (Serviço Florestal Brasileiro), 2018. Inventário Florestal Nacional. Rio de Janeiro. Principais Resultados. Ministério do Meio Ambiente, Brasilia, Brazil. https://www.gov.br/florestal/pt-br/centrais-de-conteudo/publicacoes/relatorios/relatorios-ifn/IFNRJprincipaisresultados.pdf.
Soriano-Disla, J.M., Janik, L.J., Viscarra Rossel, R.A., Macdonald, L.M., McLaughlin, M. J., 2014. The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. Appl. Spectrosc. Rev. 49, 139–186. https://doi.org/10.1080/05704928.2013.811081.
Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010. Visible and near infrared spectroscopy in soil science. In: Sparks, D.L. (Ed.), Advances in Agronomy, Advances in Agronomy, 107, pp. 163–215. https://doi.org/10.1016/s0065-2113(10)07005-7.
Stevens, A., Ramirez-Lopez, L., 2025. An introduction to the prospectr package. An Introduction to the Prospectr package. R package Version 0.2.7. https://cran.r-project.org/web/packages/prospectr/vignettes/prospectr.html.
Terra, F.S., Demattê, J.A.M., Viscarra Rossel, R.A., 2015. Spectral libraries for quantitative analyses of tropical Brazilian soils: comparing vis–NIR and mid-IR reflectance data. Geoderma 255-256, 81–93. https://doi.org/10.1016/j.geoderma.2015.04.017.
Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. B. (Methodol.) 58, 267–288. http://www.jstor.org/stable/2346178.
Vasques, G.M., Grunwald, S., Harris, W.G., 2010. Spectroscopic models of soil organic carbon in Florida, USA. J. Env. Qual 39, 923–934. https://doi.org/10.2134/jeq2009.0314.
Vasques, G.M., Balieiro, F.C., Silveira Filho, T.B., Magalhães, M.A.F., Dart, R.O., Martins, A.M.M., Andrade, B.C., Pedreira, J.P.N.C., Prado, R.B., 2025. Soil carbon stock maps for the state of Rio de Janeiro: in support of carbon sequestration and offset opportunities (Org.). In: Silveira Filho, T.B., Balieiro, F.C. (Eds.), Overview of Soil Carbon Stocks in Rio de Janeiro: An Estrategic Environmental Asset. INEA, Rio de Janeiro Brazil, pp. 24–33. https://www.alice.cnptia.embrapa.br/alice/bitstream/doc/1181256/1/Soil-carbon-stock-maps-for-the-state-of-Rio-de-Janeiro-2025.pdf.
Viscarra Rossel, R.A., Lark, R.M., 2009. Improved analysis and modelling of soil diffuse reflectance spectra using wavelets. Eur. J. Soil. Sci 60, 453–464. https://doi.org/10.1111/j.1365-2389.2009.01121.x.
Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131, 59–75. https://doi.org/10.1016/j.geoderma.2005.03.007.
Viscarra Rossel, R.A., Bui, E.N., Caritat, P., McKenzie, N.J., 2010. Mapping iron oxides and the color of Australian soil using visible–near-infrared reflectance spectra. J. Geophys. Res.:. Earth. Surf. 115 (F04031). https://doi.org/10.1029/2009JF001645.
Viscarra Rossel, R.A., Chappell, A., Caritat, P., McKenzie, N.J., 2011. On the soil information content of visible–near infrared reflectance spectra. Eur. J. Soil. Sci 62, 442–453. https://doi.org/10.1111/j.1365-2389.2011.01372.x.
Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Chabrillat, S., Demattê, J.A.M., Ge, Y., Gomez, C., Guerrero, C., Peng, Y., Ramirez-Lopez, L., Shi, Z., Stenberg, B., Webster, R., Winowiecki, L., Shen, Z., 2022. Diffuse reflectance spectroscopy for estimating soil properties: a technology for the 21st century. Eur. J. Soil. Sci 73 (e13271). https://doi.org/10.1111/ejss.13271.
Viscarra Rossel, R.A., Shen, Z., Lopez, L.R., Behrens, T., Shi, Z., Wetterlind, J., Sudduth, K.A., Stenberg, B., Guerrero, C., Gholizadeh, A., Ben-Dor, E., Luce, M.S., Orellano, C., 2024. An imperative for soil spectroscopic modelling is to think global but fit local with transfer learning. Earth-Sci. Rev. 254, 104797. https://doi.org/10.1016/j.earscirev.2024.104797.
Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemom. Intell. Lab. Syst. 58, 109–130. https://doi.org/10.1016/s0169-7439(01)00155-1.
Zuo, Y.J., Serfling, R., 2000. General notions of statistical depth function. Ann. Stat. 28, 461–482. https://doi.org/10.1214/aos/1016218226.