scientific reports

Check for updates

OPEN

Characterization and detailed mapping of C by spectral sensor for soils of the Western Plateau of São Paulo

Kathleen Fernandes[1], José Marques Júnior[1], Adriana Aparecida Ribon[2], Gabriela Mourão de Almeida[1], Mara Regina Moitinho[1]✉, Denise de Lima Dias Delarica[1], Angélica Santos Rabelo de Souza Bahia[1] & Dener Márcio da Silva Oliveira[3]

Soil mineralogy and texture are directly related to soil carbon due to the physical properties of the clay surface. Traditional techniques for quantifying carbon in soil are time-consuming and expensive, making large-scale quantification for mapping unfeasible. The alternative is the use of soil sensors, such as diffuse reflectance spectroscopy (DRS), an economical, fast, and accurate technique for predicting carbon stocks. In this sense, this study aimed to (a) investigate the relationship of C with different soil mineralogical, chemical, and physical attributes for different geological and geomorphological compartments; (b) understand which spectral bands are most important for estimating C content; (c) estimate C content from diffuse reflectance spectroscopy using different mathematical techniques and indicate which one is the best for tropical soil conditions; and (d) map C contents in detail. The study area was the Western Plateau of São Paulo (WPSP), which covers approximately 13 million hectares (~ 48% of the State of São Paulo, Brazil). A total of 265 samples were collected in this area. The attributes clay, silt, sand, crystalline and non-crystalline iron, base saturation, soil density, total pore volume, total C, C stock, kaolinite/(kaolinite + gibbsite) and hematite/(hematite + goethite), hematite and goethite contents, and spectral curves were evaluated. The spectra were recorded at 0.5-nm intervals, with an integration time of 2.43 nm s^{-1} over the 350 to 2500-nm range (350–800 nm—visible—VIS and 801–2500 nm—near-infrared—NIR). The data were subjected to descriptive statistics, Spearman correlation, stepwise analysis, and cluster grouping for characterization purposes; partial least squares regression (PLSR) and random forest (RF) for estimation purposes; and geostatistics analysis for creation of spatial maps. Our results indicate that the highest C contents are associated with more clayey soils, oxidic mineralogy, higher total pore volume, and lower soil density in highly dissected basalt compartments. The random forest algorithm associated with the Vis–NIR spectral range is more efficient for estimating and mapping C contents. This suggests that integrating diffuse reflectance spectroscopy with machine learning techniques holds promise for shaping public policies related to land use, mitigating CO$_2$ emissions, and facilitating the implementation of carbon credit policies in a rapid and economically efficient manner.

Keywords Chemometrics, Machine learning, Global warming, Soil mineralogy

World soils contain around four times as much carbon (C) as the vegetation and three times as much as the atmosphere[1,2]. Globally, soil organic C (SOC) stocks are currently estimated to be 1400 ± 150 petagrams of carbon (Pg C) at 1 m in depth and 2060 220 Pg C to 2 m at depth[3]. Additionally, SOC improves soil health and the delivery of other related ecosystem services[4,5], with undeniable benefits on crop yields, boosting or maintaining the production of food, feed, fiber, and energy. Accordingly, any change in the soil C reservoir would significantly impact both world food security and global climate change[6,7]. In this sense, assessing and monitoring

[1]School of Agricultural and Veterinary Sciences, São Paulo State University (FCAV-UNESP), Via de Acesso Prof. Paulo Donato Castellane, s/n, Jaboticabal, São Paulo, Brazil. [2]Goiás State University (UEG), West Campus, University Unit of Palmeiras de Goiás, S-7, s/n Setor Sul, Palmeiras de Goiás, Goiás, Brazil. [3]Federal University of Viçosa (UFV), Campus Florestal. Rodovia LMG 818, km 06, Florestal, Minas Gerais, Brazil. ✉email: maramoitinho@gmail.com

SOC is mandatory[1,8]. However, the conventional laboratory-based analytical methods for SOC accounting are expensive and time-consuming[9,10]. The development of accurate and cost-efficient alternative techniques for SOC quantification is a major concern in the climate policy[11].

Diffuse reflectance spectroscopy (DRS) is a technique that records the absorption and dispersion of light processes, produced on the surface of soil particles[12]. It is a fast, non-polluting technique of lower cost and has already been used to characterize different attributes[13–16]. Its use for estimating soil C content[17–20] may allow the detailed mapping of this variable on large scales and the understanding of its spatial variation in landscapes[8,21].

In addition to minerals in the soil governing numerous properties directly related to crop productivity[22–24], soil mineralogy plays a large role in directly regulating the size and stability of C stocks[25]. Some studies have indicated a direct relationship between oxides and C flow in the soil of small areas[15,18,26]. However, the existing relationships between mineralogy and C in large areas, with higher geological and geomorphological variability, are still little studied. Understanding the relationship between soil attributes in detail is the best alternative for management practices and soil planning[27].

Multivariate analyses, such as partial least squares regression (PLSR)[28], and machine learning algorithms, such as random forest (RF)[29], have been widely used in C estimation using DRS data[18–21,30] Bellon-Maurel and McBratney[17] provided a detailed literature review about DRS and C estimates. According to the authors, the future proposal is to optimize these studies spatially for mapping purposes and associate the estimates with robust mathematical models.

Soil is a system that is constantly changing. Therefore, accurate predictions of the consequences of human activities on terrestrial ecosystems may be difficult[31]. The diffusion of new technologies and agricultural research developed in the coming decades will determine the possibility of mitigating resources and adapting to climate change to ensure continuity in food production[32]. Moreover, there is always a shortage of data, inventories, detailed C maps for large-scale projects[21,33], and the best mathematical techniques for these studies[30,34], mainly in Brazil.

Studies of wide magnitude, which present not only C estimates but also its relationship with other soil attributes and their spatial variability on maps are essential for agriculture and the sustainability of terrestrial ecosystems. In this context, this study aimed to (a) investigate the relationship of C with different soil mineralogical, chemical, and physical attributes for different geological and geomorphological compartments; (b) understand which spectral bands are most important for estimating C content; (c) estimate C content from diffuse reflectance spectroscopy using different mathematical techniques and indicate which one is the best for tropical soil conditions; and (d) map C contents of the Western Plateau of São Paulo in detail.

Material and methods

Location and characterization of the area

The study area was the Western Plateau of São Paulo (WPSP), which covers approximately 13 million hectares (~48% of the State of São Paulo, Brazil) (Fig. 1). The geological outline is mainly characterized by sandy, clayey, and gravel sediments, volcanic rocks with basic composition, and sedimentary sequences, mainly psamitic, which may include pyroclastic sequences (Fig. 1a). Approximately 2 million hectares of WPSP are represented by basalt (15.5%), 7.4 million hectares by the Peixe River Valley Formation (57.1%), and 3.6 million hectares by other sedimentary formations (27.5%).

The geological formations of WPSP are divided into two groups: (i) Caiuá group, composed of Santo Anastácio and Paraná River formations in the State of São Paulo, corresponding to deposits of sand sheets of dry climate, accumulated in extensive and monotonous desert plains, marginal to the large sand sea dune complexes (Caiuá Desert), extending to the northern region of the State of Paraná; and (ii) Bauru group, composed of Uberaba, Peixe River Valley, Araçatuba, São José do Rio Preto, Presidente Prudente, and Marília formations, which include Taiúva analcimites and volcanic rocks locally interspersed in the sequence[35,36].

Soils with the highest occurrence are classified as Argissolo Vermelho-Amarelo (Ultisol), Latossolo (Oxisol), Latossolo férrico (Oxisol), Neossolo Litólico (Entisol), Nitossolo Vermelho (Nitisol), and Gleissolo Háplico (Gleysol) (Fig. 1b). The geomorphology of the area is shown in Fig. 1d. Moderately dissected areas are predominant in the region. The dissection level expresses the soil evolution in the landscape (Fig. 1b), which is associated with pedogenesis (soil formation rate) and geomorphogenesis (landscape evolution rate). The lowest hematite (Hm) content, highest amorphous goethite (Gt) contents[37], and highest kaolinite (Kt) proportion are observed in the central portion of WPSP, where sandstones prevail[38].

The mineralogical characterization at the detail level can infer numerous other soil characteristics. Silva et al.[37] and Fernandes et al.[38] pointed out that iron oxides and minerals Kt and Gb accurately reflect their respective formation environments and the variability imposed by geological and geomorphological material. Soils undergo a reduction process in humid environments in highly dissected compartments in the most drained regions of the landscape, favoring Gt formation. On the contrary, sandstone compartments, rich in silica, favor Kt formation.

The tropical climate with a dry winter season prevails in the north and northwest of WPSP (C2rA′a′), while the humid temperate climate with a hot summer prevails in the south (B4rB′4a). In addition, the climate in the east and southeast consists of a humid temperate climate with a dry winter and hot summer (B2rB′3a). The climate in these regions can be subclassified into four more variations, according to the Thornthwaite classification. The natural vegetation in the area consisted of Atlantic Forest in the west and Cerrado in the east and southwest of WPSP, with the most representative current land use being sugarcane (*Saccharum* spp.), citrus, and pasture.

A total of 265 soil samples georeferenced in WPSP were collected to study the spatial variability of mineralogical attributes at a depth of 0.00–0.20 m. The minimum spacing between samples was 10 km and the maximum spacing was 60 km (Fig. 1d). The samples were collected along the highways of the State of São Paulo in representative locations that have suffered minimal anthropogenic interference. The sampling plan was prepared

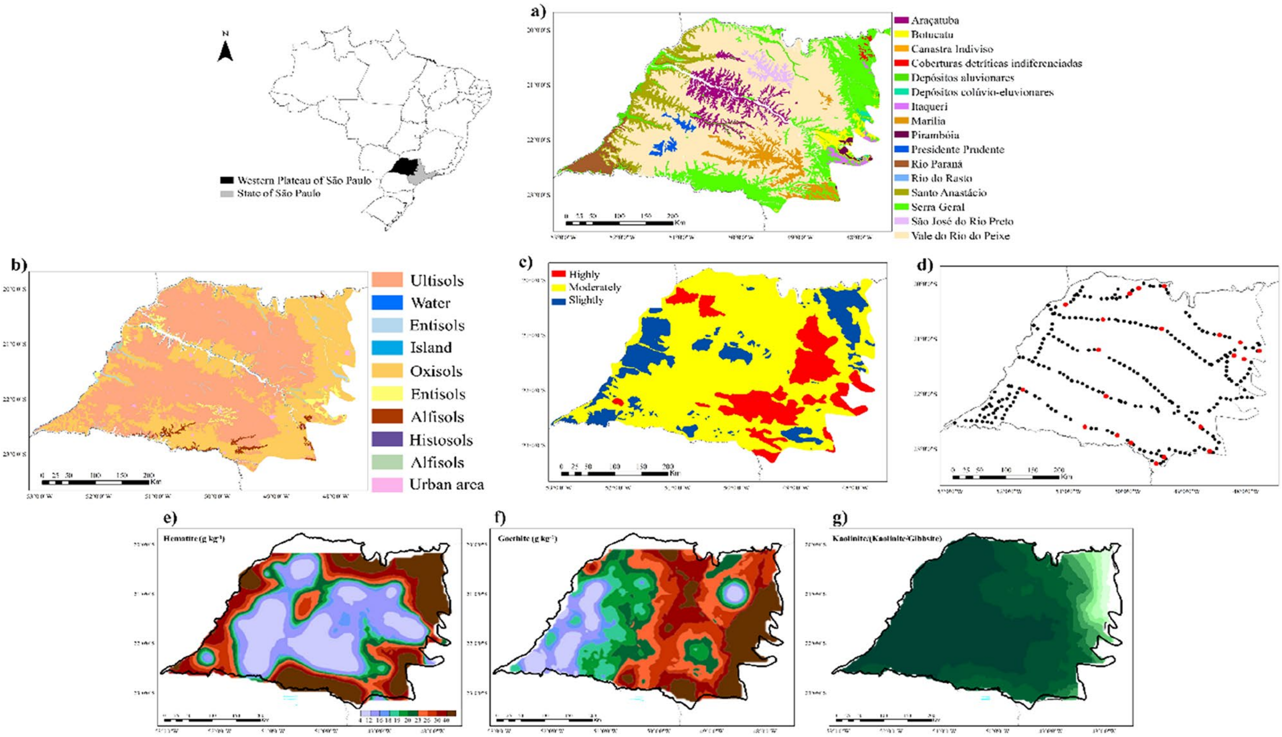

Figure 1. (**a**) Updated geological map of the Western Plateau of São Paulo. (**b**) Pedological map of the Western Plateau of São Paulo. Source: Agronomic Institute of Campinas. (**c**) Dissection map of the Western Plateau of São Paulo. (**d**) Sample planning of the collected points. (**e**) Hematite map made by X-ray diffraction (Hm, g kg^{-1}). Source: Extracted from Silva et al.[37] (**f**) Goethite map made by X-ray diffraction (Gt, g kg^{-1}). Source: Extracted from Silva et al.[37]. (**g**) Kaolinite/(kaolinite/gibbsite) ratio [Kt/(Kt + Gb)] map made by X-ray diffraction. Source: Extracted from Fernandes et al.[38].

based on the file of State highways supplied by the Department of Roads and Highways using the ET GeoWizards tool in the software ArcView 9.3. The geographic coordinate information for each point was previously defined and inserted in a GPS navigation device with an accuracy of 3–5 m. Sample collection was guided by real-time navigation using GPS and laptops.

Points simulating the stratified sampling (in red in Fig. 1d) were selected after identifying the patterns of variability in soil attributes and identifying homogeneous areas[39]. The choice of the sampled points considered previous experiences within WPSP, using geostatistical techniques[40,41], proposing a representative distribution of the studied area.

Analyses

Soil particle size and chemical analysis

Particle size analyses were performed for all WPSP points. The pipette method with a 0.1 mol L^{-1} NaOH solution as a chemical dispersant and mechanical stirring at a low rotation for 16 h was used, as recommended by Teixeira[42]. Calcium, magnesium, and potassium contents were extracted using the ion-exchange resin procedure[43]. Base saturation (V) was given by the equation V = (Mg + Ca + K)/CEC.

Mineralogical analyses

Iron contents related to the totality of pedogenetic irons extracted by dithionite-citrate-bicarbonate (Fed) were determined following the procedure of Mehra and Jackson[44] for all grid points and the stratified samples. The contents of iron extracted by ammonium oxalate (Feo) relative to low-crystallinity pedogenetic iron oxides were determined following the methodology mentioned by Camargo et al.[45] adapted from Schwertmann.[46].

Minerals were quantified for all points on the total grid and stratified samples. Clay for the X-ray diffraction (XRD) analysis was separated from the soil sample by the centrifugation method[47]. The clay fraction was subjected to the elimination of iron oxides by the dithionite-citrate-bicarbonate (DCB) method to characterize kaolinite (Kt) and gibbsite (Gb), according to Mehra and Jackson[48], and sieved in a 0.10-mm opening mesh. XRD characterized the minerals of the clay fraction Kt and Gb in sheets made with material without orientation (powder). Goethite (Gt) and hematite (Hm) were characterized after treating the clay fraction using 5 mol L^{-1} NaOH (1 g clay 100 mL^{-1} solution) for their concentration, according to the method of Norrish and Taylor[49], modified by Kämpf and Schwertmann[50].

XRD was performed with the samples prepared by the powder method using an HGZ apparatus equipped with cobalt cathode and iron filter and K radiation (20 mA, 30 kV) for Hm and Gt diffraction and copper cathode with nickel filter for Kt and Gb diffraction. The scanning speed was 1°2θ min^{-1} with an amplitude from 23°

to 49°. The reflexes of Kt (001), Gb (002), Hm (012 and 110), and Gt (110 and 111) were used for mineralogical evaluation.

Diffuse reflectance spectroscopy (DRS)

Diffuse reflectance spectra were obtained using approximately 1 g of air-dried fine soil ground in an agate mortar until obtaining constant color, and the content was placed in sample holders with a 16-mm diameter cylindrical space. The soil subsamples (1 g) used for the DRS analysis came from the samples corresponding to 250 g of soil collected at each of the 265 points georeferenced in WPSP.

Reflectance values were determined in a Lambda 950 UV/Vis/NIR spectrophotometer coupled to an integrating sphere of 150 mm in diameter. The spectra were recorded at 0.5-nm intervals, with an integration time of 2.43 nm s^{-1} over the 350 to 2500-nm range (350–800 nm—visible—VIS and 801–2500 nm—near-infrared—NIR).

The pre-processing technique used to remove noise from the raw curves was the standard normal variable (SVN). This technique is used to remove interference due to light scattering, helping to correct any curvilinear and linear trends in the baseline of the original spectra[14].

Carbon content and stock

The total soil carbon content (C%) was determined for all points of WPSP. The analysis was carried out by dry combustion using a LECO CN-2000[9]. The soil samples (air-dried fine soil) were sieved on a 100-mesh sieve and oxidized at high temperatures (oven at 1350 °C) using 2.8 ultrapure oxygen.

The C stock was analyzed only for stratified samples. Undisturbed soil samples were collected with a volumetric ring (with known volume) at a depth of 0.00–0.05 m to determine the following physical attributes: porosity, soil density (Ds), and soil moisture. The respective masses and soil moisture in the ring were determined for Ds calculation. The density of each layer was calculated according to the equation:

$$\text{Ds}\left(\text{g cm}^{-3}\right) = (\text{dry soil mass, g})/\left(\text{cylinder volume, cm}^{3}\right)$$

The total contents of C stock were determined using cross-calibration curves and based on the weight of the analyzed sample. C stocks (in Mg ha^{-1}) were calculated for each soil sample according to the equation described below:

$$\text{C stock} = \text{C (\%)} \times \text{Ds}\left(\text{g cm}^{-3}\right) \times \text{Thickness (cm)}$$

Statistical analysis

The data were subjected to descriptive statistics after the completion of laboratory analysis. The mean, maximum, minimum, standard deviation, and coefficient of variation were calculated. Moreover, simple correlation coefficients (Spearman) were calculated between soil attributes. All wavelengths of the spectral curves were submitted to the stepwise regression gradually and interactively (forward–backward) based on the Akaike information criterion (AIC). Subsequently, the most relevant wavelengths for estimating C content were selected, as follows. These wavelengths were subjected to cluster grouping analysis.

Two C analyses were used to estimate C contents: partial least squares regression (PLSR) and machine learning (ML). PLSR uses chemometric calibrations and validations by a cross-validation procedure. Reflectance measurements were converted into absorbance [Log10 (1/Reflectance)] for data processing in the software. The software ParleS®[28] was used to determine the relationship between the entire spectral curve and the studied attributes. The database was subdivided into a set of calibrations (70%) and a set of predictions (30%) for analysis. The "train_test_split" technique was used to split the data into calibration and validation sets. This is a function in the "model_selection" module of the popular machine learning library scikit-learn. This function is used to perform the train test split procedures, which splits a dataset into two subsets: a training set and a test set.

The performance of each model was evaluated using the coefficient of determination (R^2), mean absolute error (MAE), and the root mean square error (RMSE). These metrics were evaluated both in calibration and in prediction. The ML algorithm was the random forest (RF) regression[51]. RF is a non-parametric technique developed to improve the prediction of the classification and regression tree models, as it uses fully grown decision trees and reduces the error and variation[52]. It consists of the combination of several trees, which are generated from an input variable sampled at random. All trees are the same size. However, the subdivisions of the trees are based on a subset randomly sampled from the total database, and the final result of RF consists of the mean of the results of all trees[53]. RF has been one of the most used ML algorithms for presenting high performance in the prediction of soil attributes[54].

The performance of prediction models was evaluated in a set of independent and unprecedented data. In this sense, the database was subdivided into two different sets randomly separated, remaining 50% for training and 50% for testing. This division was due to the high variability of the data. The performance of each model was evaluated using R^2 and RMSE in the training and testing stages. The package scikit-learn was used to run the algorithm[55].

The geostatistical analysis was used to characterize the spatial variability pattern of the observed values[56]. Semivariograms relating the distance vector to the semivariance were constructed and models were adjusted. The values were interpolated by ordinary kriging for the construction of spatial distribution maps. The semivariograms were generated and selected using the R language, which already provides the best adjustment in the geostat package.

Results and discussion

Soil attributes across the Western Plateau of São Paulo

Soils originating from sandstone had a mean sand of 82%, while those from basalt were more clayey, with a mean of 53% clay (Table 1). The maximum and minimum clay values for basalt soils reached 67 and 27%, respectively. This higher range between minimum and maximum shows the variation of the attribute for the compartment. The same can be observed by the high coefficients of variation, according to the classification of Warrick and Nielson[57]. Sandstone soils have a maximum of 90% and a minimum of 73% sand and, therefore, less variability compared to basalt soils. The CV for sand content was classified as low for the compartment.

The mean clay values in the geomorphological compartments increased from slightly (23% clay) to highly dissected compartments (34% clay). The means of silt and sand tended to increase from the highly to the slightly dissected compartments (Table 1). The maximum and minimum values for clay and silt means were higher for the low dissected compartment than the others. It shows that soils of highly dissected compartments tend to be more clayey and those of slightly dissected compartments tend to be sandier.

Slightly dissected environments would be those flatter ones, favoring higher water infiltration, better drainage, higher leaching, more uniform solar radiation, and other characteristics that condition higher soil weathering (higher pedogenesis rates). Therefore, higher mean values of clay and lower mean values of sand were expected. However, considering that highly dissected compartments have concave areas of high intensity and a depositional behavior[58], finer materials, such as clay and organic material, may have accumulated in this compartment. Cunha et al.[59] and Silva et al.[37] observed similar results.

Basalt compartment soils have higher mean contents of Fed (52 g kg^{-1}) than those from sandstone (12 g kg^{-1}) (Table 1). The same was observed for the Feo content, in which sandstone compartment soils presented a mean of approximately 1 g kg^{-1}, while basalt soils had a mean of 3 g kg^{-1}. The iron content by geomorphological

	Clay	Silt	Sand	Fed	Feo	V	Ds	TPV	C	Cs	Kt/	Gt/	Gt	Hm
	%			g kg^{-1}		%	g cm^{-3}	%	g kg^{-1}	Mg kg^{-1}	(Kt + Gb)	(Gt + Hm)	g kg^{-1}	
Sandstone compartment soils														
Mean	14.55	3.18	82.27	12.58	0.87	65.04	1.51	36.09	10.74	80.75	0.89	0.29	5.12	12.15
Min	7.49	0.50	73.40	5.78	0.47	36.55	1.31	30.21	5.32	42.03	0.41	0.08	1.31	3.56
Max	22.50	11.34	90.86	18.21	1.94	78.87	1.69	40.87	18.71	147.61	1.00	0.60	8.59	19.88
SD	4.74	3.13	5.28	4.05	0.37	11.50	0.13	3.90	4.34	33.36	0.18	0.14	2.41	5.18
CV	32.59	98.44	6.41	32.21	42.75	17.68	8.64	10.82	40.39	41.32	19.89	47.17	47.07	42.63
Basalt compartment soils														
Mean	53.10	9.46	37.44	52.22	3.34	70.25	1.39	41.78	20.82	144.34	0.46	0.22	15.63	55.84
Min	27.50	1.55	23.89	29.95	1.50	43.64	1.27	32.50	14.85	102.02	0.29	0.03	2.88	23.82
Max	67.47	20.37	70.95	64.33	6.08	84.55	1.69	47.67	25.95	189.12	0.71	0.50	26.06	73.65
SD	15.79	6.92	20.06	12.61	1.39	12.39	0.14	5.04	3.78	29.14	0.15	0.13	7.14	18.57
CV	29.74	73.10	53.60	24.14	41.73	17.64	10.26	12.07	18.15	20.19	33.36	61.03	45.70	33.26
Highly dissected compartment soils														
Mean	34.36	3.77	61.87	29.32	1.72	71.17	1.39	40.62	19.76	137.55	0.71	0.29	10.25	30.25
Min	19.98	0.60	25.30	8.49	0.56	67.36	1.29	36.99	15.49	110.72	0.38	0.03	2.88	7.73
Max	65.00	9.70	76.92	56.89	3.57	74.12	1.46	45.64	25.95	189.12	1.00	0.50	26.06	73.65
SD	20.86	4.09	24.71	21.45	1.37	3.21	0.08	3.68	4.55	35.23	0.25	0.19	10.77	29.67
CV	60.72	108.29	39.94	73.14	79.69	4.51	5.41	9.06	23.00	25.61	35.67	65.72	105.13	98.07
Moderately dissected compartment soils														
Mean	28.07	5.36	66.57	27.04	1.86	65.71	1.49	37.42	13.02	95.65	0.74	0.23	8.27	28.56
Min	10.00	0.50	23.89	8.60	0.47	36.55	1.27	30.21	5.32	42.03	0.29	0.08	1.31	5.97
Max	67.47	18.64	89.36	64.33	6.08	78.87	1.69	47.67	23.97	172.33	1.00	0.39	18.91	69.02
SD	22.04	5.05	25.26	20.82	1.62	12.00	0.16	5.64	6.31	44.95	0.30	0.11	5.28	23.50
CV	78.52	94.31	37.95	77.00	86.93	18.26	10.64	15.07	48.51	46.99	40.46	45.25	63.77	82.27
Slightly dissected compartment soils														
Mean	23.30	8.26	68.44	23.70	1.37	67.41	1.44	38.58	14.18	100.09	0.77	0.40	10.54	22.46
Min	7.49	1.65	27.21	5.78	0.52	45.61	1.31	35.43	8.48	61.61	0.61	0.26	2.71	3.56
Max	52.42	20.37	90.86	58.57	2.92	84.55	1.57	42.05	21.97	143.68	1.00	0.60	22.95	58.94
SD	25.25	10.51	35.75	30.20	1.35	19.88	0.13	3.32	6.98	41.27	0.21	0.18	10.87	31.60
CV	108.33	127.24	52.24	127.43	98.76	29.50	9.10	8.61	49.23	41.23	26.99	45.16	103.16	140.67

Table 1. Descriptive statistics for soil attributes from different geological and geomorphological compartments in agricultural areas and native forest in the Western Plateau of São Paulo. *Min* minimum, *Max* maximum, *SD* standard deviation, *CV* coefficient of variation, *Fed* crystalline iron, *Feo* non-crystalline iron, *V* base saturation, *Ds* soil density, *TPV* total pore volume, *C* carbon, *Cs* carbon stock, *Kt* kaolinite, *Gb* gibbsite, *Gt* goethite, *Hm* hematite.

compartment showed that the slightly dissected compartment had the lowest means of Fed (24 g kg^{-1}) and Feo (1.37 g kg^{-1}). Fed content had the same behavior as the clay content, that is, lower values in the slightly dissected compartment and higher values in the highly dissected compartment.

Basalt compartment soils showed a higher V (70%) than soils originating from sandstone (65%) (Table 1). WPSP is an intensely cultivated region and soil liming/fertilization and management practices may have influenced these results. Also, sandstone rocks have higher particle sizes and, consequently, the weathering process is slower in these rocks than in basalt rocks. The mean of V for geomorphological compartments reached 71%, while soils from the slightly dissected environment showed a mean of 67%. The moderately dissected compartment was intermediate compared to the others (65%). The differences are small for these compartments.

The mean value of soil density (Ds) was higher for sandstone soils (1.52 g cm^{-3}) than for basalt soils (1.39 g cm^{-3}) (Table 1). Thus, the data analysis by compartments for large areas may present results similar to those observed in the literature. In this case, the shape of the mineral in kaolinitic soils, such as sandstones, may favor higher Ds values, while the shape of these minerals for basalt soils, with higher iron oxide content, may contribute to better arrangement of particles and lower Ds[60]. Ds presented intermediate means for the moderately dissected compartment (1.49 g cm^{-3}), increasing from highly to slightly dissected (1.39–1.44 g cm^{-3}, respectively).

The means of total pore volume (TPV) showed an inverse behavior when compared to the means of Ds (Table 1). TPV is higher for basalt compartment soils (41%) and lower for sandstone compartment soils (36%). The means for geomorphological compartments had little difference, i.e., 40, 37, and 38% for highly, moderately, and slightly dissected compartments, respectively. It indicates that the geomorphological compartment did not capture TPV variability well.

The total C contents were higher for areas of basalt and highly dissected compartments than the other compartments. The mean for the basalt compartment (20.82 g kg^{-1}) was almost double the mean observed for the sandstone compartment (10.74 g kg^{-1}). Bahia et al.[18] observed C values similar to those found in this research when studying smaller WPSP areas. All mean attribute values are similar to those found in the literature for different sandstone and basalt geological compartments[18,60,61].

The basalt compartment presented the highest C stock (Cs) (144 Mg ha^{-1}) than the sandstone compartment. The same behavior was found for highly (137 Mg ha^{-1}) and slightly dissected compartments (100 Mg ha^{-1}), matching the higher C and clay contents. The highest Cs values are related to compartments with predominant oxidic mineralogy (Table 1), as verified for the entire WPSP extension.

Sandstone compartment soils are kaolinitic and basalt compartment soils are oxidic, showing the highest mean values of Gt (15 g kg^{-1}) and Hm (55 g kg^{-1}). It is also observed in the mean value for the Gt/(Gt + Hm) ratio of 0.22 for soils originating from basalt and the mean of 0.89 for the Kt/(Kt + Gb) ratio for sandstone soils (Table 1). According to Curi and Franzmeier[62], soils originating from basalt present higher iron contents than those originating from sandstone, favoring the formation of iron oxides to the detriment of other minerals. Mineralogy presented means with small differences for geomorphological compartments. The highest difference was found for the Gt/(Gt + Hm) ratio of 0.40 in the slightly dissected compartment, which was higher than the others. Thus, this compartment has a predominance of oxidic mineralogy, while the others (highly and moderately dissected) have a predominance of kaolinitic mineralogy.

Correlation between C contents and soil attributes

Figure 2 shows a correlation matrix with all the studied attributes and geological and geomorphological compartments. A direct correlation can be observed between sand content and the sandstone compartment and between clay content and the basalt compartment, as found in the descriptive statistics. Sandstone soils are sandier and basalt soils are more clayey. Geomorphological compartments did not show good correlations with geological compartments. A directly proportional correlation was found between clay content and the highly dissected compartment and an inverse correlation with sand content. An inverse behavior was observed for sand content.

Fed and Feo contents showed a high and inverse correlation with the sandstone compartment (− 0.83) and a direct correlation with the basalt compartment (Fig. 2). Regarding the geomorphology, Fed and Feo contents showed low correlation coefficients, with the highest values of − 0.24 and − 0.22 for soils of the slightly dissected compartment, respectively. The attributes V, TPV, C, Cs, Gt, and Hm showed a direct correlation with the basalt compartment and an inverse correlation with the sandstone compartment, while Ds and the Kt/(Kt + Gb) and Gt/(Gt + Hm) ratios showed negative correlations with the sandstone compartment and positive correlation with the basalt compartment.

Considering the correlations between the attributes and C and Cs contents, the highest values were found between clay content and soil mineralogy. The higher the clay content and predominance of soil oxides, the higher the C and Cs contents. The opposite was observed for sandier soils, with a predominance of kaolinitic mineralogy, as the correlation between these attributes and C and Cs was high but negative. Therefore, areas with higher means of clay content are potential C reservoirs, as proposed by Mendes[63]. Soil CO_2 emission can be described by the emission and C stock decay coefficients. Thus, the potentials for higher or lower emissions are intrinsic to the soil and closely related to formation processes and factors, expressed by covariate attributes, such as texture and mineralogy[15,26,64].

Carbon stocks vary according to the studied soil type and depth, and although these correlations are poorly studied, the stock is an important indicator of environmental services[65]. Assad et al.[66] observed that soil properties gain prominence in the influence of Cs on local work scales. Saiz et al.[67] pointed out that soil texture was a determinant factor for Cs variations.

The descriptive statistics and all the studied attributes show that basalt soils and slightly dissected compartments are more weathered than sandstone soils and highly dissected compartments. The moderately dissected compartment presented intermediate means for highly and slightly dissected compartments for all attributes,

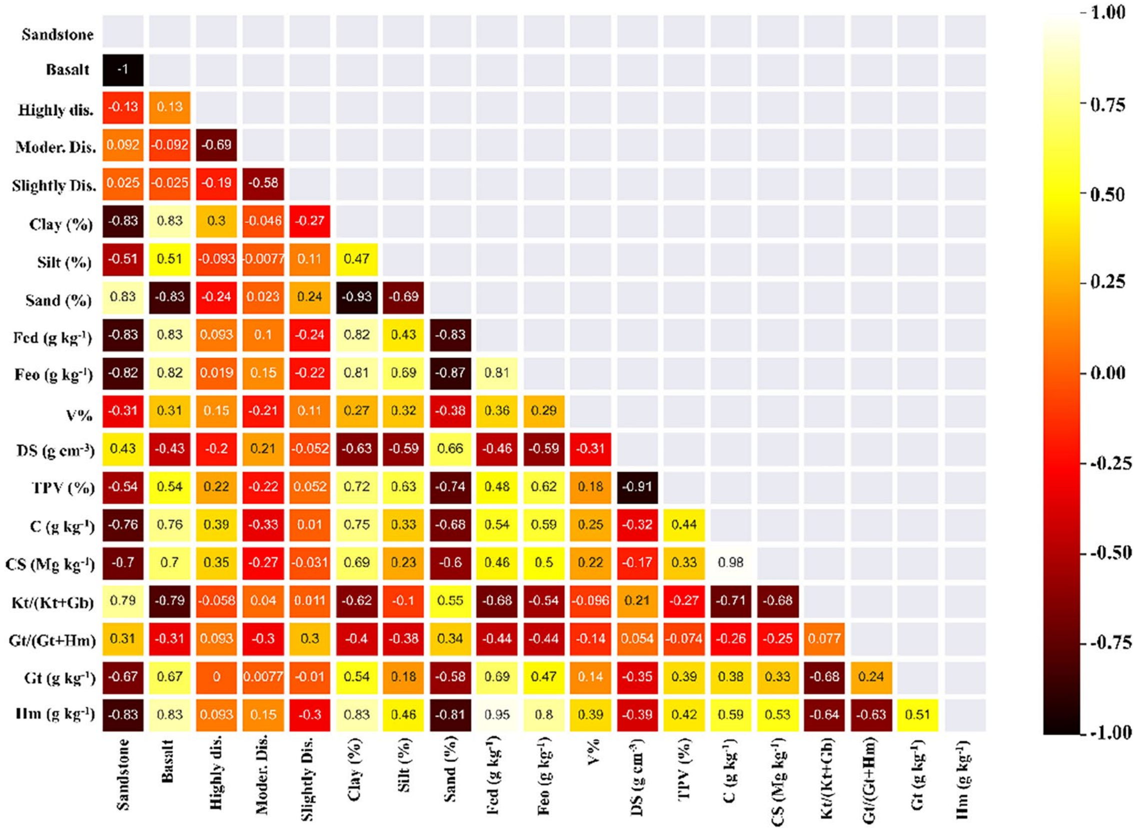

Figure 2. Spearman correlation matrix for attributes at a depth of 0–0.20 m and different geological and geomorphological compartments of the Western Plateau of São Paulo. *Fed* crystalline iron, *Feo* non-crystalline iron, *V* base saturation, *Ds* soil density, *TPV* total pore volume, *C* carbon, *Cs* carbon stock, *Kt* kaolinite, *Gb* gibbsite, *Gt* goethite, *Hm* hematite.

not allowing for a more specific differentiation. All intermediate landscape forms were included in this group. Therefore, the attributes are expected to present higher variation than the other compartments (Table 1).

The intensity of landscape shapes, indicated by the dissection level, follows a pattern imposed by the structural control of the parent material, indicating a pedogenesis (soil formation) and geomorphogenesis variation (landscape carving). The highly dissected compartment shows concave areas of high intensity, indicating higher dissections in the region and, consequently, higher geomorphogenesis, while the slightly dissected compartment encompasses more preserved regions, with more flattened surfaces and higher pedogenesis rates than geomorphogenesis.

Vasconcelos et al.[68] studied the pedo-geomorphogenesis evolution model of Serra da Canastra and observed that the dynamics of the relief developed from the process of cutting the land (higher landscape dissection) causes environments of water saturation and stagnation and better conditions for drainage in flatter areas. However, environments of higher water stagnation favor the advance of hydromorphy, resulting in the formation of amorphous minerals and Kt predominance[69], while flatter and well-drained areas favor the advance of oxidation, with the domain of more crystalline minerals and higher Gb content[70].

The observations showed that highly dissected basalt compartments with higher clay content and oxidic mineralogy are more favorable for C storing. The opposite occurs for sandstone soils. These results can assist producers in making decisions regarding planting, reforestation, and practices aiming to increase Cs in the soil and reduce CO_2 emissions. Some of these practices are cited by Minasny et al.[8] The use of these practices has already been considered in C credit and monetization policies.

Spectral signature and most important bands for C estimation

Figure 3 shows the spectral signatures for the different geological and geomorphological compartments of WPSP. The curves for the sandstone soil presented higher reflectance. Sandstone soils have a higher content of clear minerals, such as orthoclase, quartz, and plagioclase, which increase the reflectance of samples[16]. On the other hand, the reflectance values for basalt samples are lower. Darker minerals in soils originating from basalt, such as iron oxides, reflect less, generating a shorter spectral curve than the curve of sandstone soils[71].

The range from 400 to 690 nm in the visible (VIS) is used to characterize the presence of the oxides Hm and Gt[14–16,35,72]. The curve in this band has a greater concavity for the basalt soil, pointing out a higher expression of oxides for soils with this parent material. The spectral behavior for basalt soils also points to a predominance of Hm relative to Gt (greater characteristic concavity). Genú et al.[73] observed similar results for Oxisols of the Serra Geral Formation.

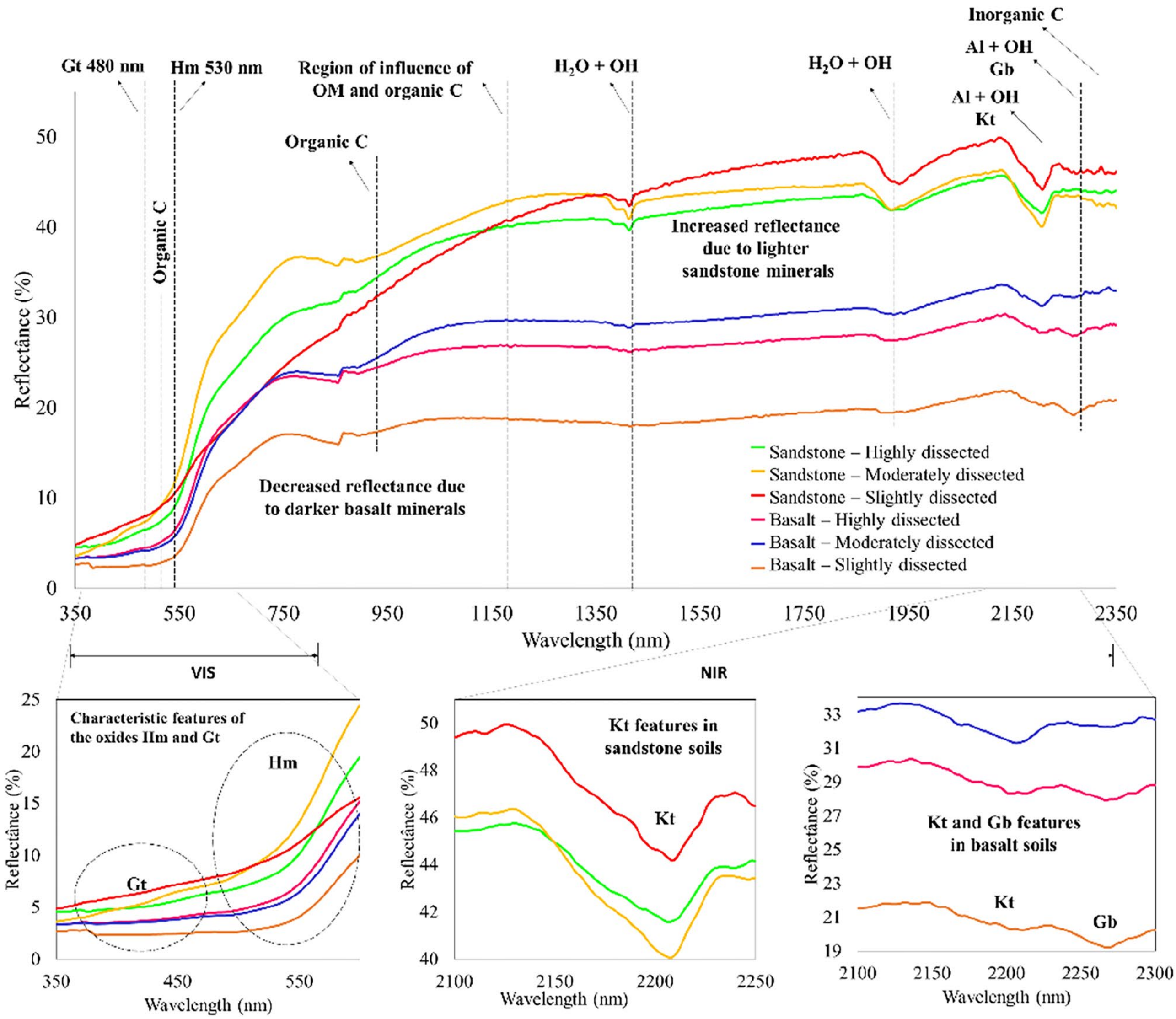

Figure 3. Spectral curves of stratified samples for sandstone and basalt in the area of the Western Plateau of São Paulo. Adapted from Fernandes et al.[38]. ***Kt* kaolinite, *Gb* gibbsite, *Gt* goethite, *Hm* hematite, *OM* organic matter, *H_2O* water molecule, *OH* hydroxyl. The ranges most frequently cited in the literature for C content are 850–900[74,80] nm and 2320–2370 nm for organic C[74,80], inorganic C, and OM[20], 410, 570, and 660 nm for organic C[81], and 476 and 808 nm[30].

The characteristic range for Kt and Gb varies from 2000–2100 to 2300–2350 nm[38,72]. The sandstone soil presented only one characteristic valley (2100–2200 nm), referring to Kt. The predominant mineral in sandstone samples, rich in quartz and silicon, was Kt. Therefore, this mineral is more abundant in the clay fraction in WPSP soils, as most of the soils in the area are of sandstone origin. Silva et al.[37] observed values of weathering indices (Ki and Kr) that classify the reference soils of the region as kaolinitic to oxidic kaolinitic.

Two features were observed for the basalt soil in the range from 2150 to 2300 nm, the first referring to Kt and the second to Gb. The valley depth of Kt was smaller regarding the curve of the sandstone soil, showing a less favorable environment for Kt formation in basalt soils. Demattê et al.[16] studied spectral curves of Brazilian biomes and observed that the characteristic features of Gb are found in more weathered soils, such as those of the Cerrado biome. On the contrary, less weathered soils, such as those of the State of Rio Grande do Norte, presented no characteristic features of Gb. Basalt has minerals more easily weathered than the minerals present in the sandstone, thus showing higher Gb content.

Stenberg and Viscarra Rossel[74] pointed out the presence of hydroxyls in the range of 1400 nm related to molecular water. A marked presence of water molecules (H–O–H) bound to minerals or impurities has been observed in the range of 1900 nm[75–77]. Dufréchou et al.[78] observed that the band of 1900 nm is not characteristic of a specific mineral but due to the clay composition. These bands were indicated in the importance values as important in the prediction of the Kt/(Kt + Gb) ratio, thus being only the influences of other minerals or the water molecule.

Moreover, Bishop et al.[79] pointed out that variations in absorbance ranges occur due to the stretching of Si, Fe, and Al oxides, cation size, and folding mode, among other crystallographic characteristics that can divide the occurrence of attributes into several spectral bands. Demattê et al.[78] observed that the relationship between the spectral behavior of soils varies according to the different soil pedogenetic processes and landscape position. Therefore, the identification of spectral bands is restricted to the characteristics and conditions of soil formation.

The ranges most frequently cited in the literature for C content are 850–900[74,80] nm and 2320–2370 nm for organic C[74,80], inorganic C, and OM[20], 410, 570, and 660 nm for organic C[81], and 476 and 808 nm[30]. Several ranges in the Vis–NIR range have the potential to predict the total, organic, or inorganic C content. Bahia et al.[18] observed that the best prediction parameters can be observed in the NIR range, as there are several ranges related to the biding of elements such as C–C, C=C, CH, and C–N over this range.

All spectral bands were inserted into the stepwise model to minimize errors. This procedure gradually and interactively selected the most important bands to predict the total C. The error measured by RMSE and MAE decreased and R^2 increased as the wavelength was inserted into the model. The first band for the Vis range to be inserted into the model was 699 nm, followed by 352, 696, 697, 503, 501, 455, 442, and 478 nm. The best model presented the following parameters: RMSE of 4.74, MAE of 3.77, and R^2 of 0.30.

The first band for the NIR range to enter the model was 2434 nm, followed by 2211, 2413, 1696, 1617, 2207, 1946, 1465, 1400, 1078, 1022, 1902, 1696, 1493, 1410, 1943, 1939, 1896, 1892, 1522, 1453, 1402, 1420, 1129, 1092, and 949 nm. The best model found for the NIR range presented an RMSE of 3.87, MAE of 3.07, and R^2 of 0.52.

The best parameters obtained for NIR can be explained according to the observations of Bahia et al.[18] as previously mentioned. More bands related to C can be found in the NIR range. Some of the bands classified as important, according to the stepwise model, were similar to those mentioned by other authors, such as 660 and 410 nm[81] and 808 and 476 nm[30]. These results indicate that some variations in the most important ranges for estimating C can be observed when comparing studies of different soils and regions around the world.

Cluster plots with heatmaps were generated from the most important bands (Fig. 4) to group the spectral bands, C contents, and geological compartments and understand their cause-and-effect relationships with the wavelengths selected by stepwise. Figure 4 shows that the highest C contents (lighter colors) were grouped with the basalt compartment for both Vis (Fig. 4A) and NIR ranges (Fig. 4B). The sandstone compartment was grouped and associated with the lowest C contents.

The most important spectral bands in the Vis and NIR ranges (Fig. 4) showed the highest C contents grouped with the lowest reflectances. C contents in the Vis range, which expresses the soil color[14,15,24], are associated with organic C, present in organic matter (OM). Demattê et al.[82] studied the influence of OM on soil reflectance and pointed out that its removal increases reflectance across the spectral curve. In other words, the lower the OM content, the lighter the soils and the higher their reflectance.

The opposite was observed for the sandstone compartment. The highest reflectances were grouped for sandstone soils with lower C contents. These facts can be observed in the spectral curves of Fig. 3. The lowest

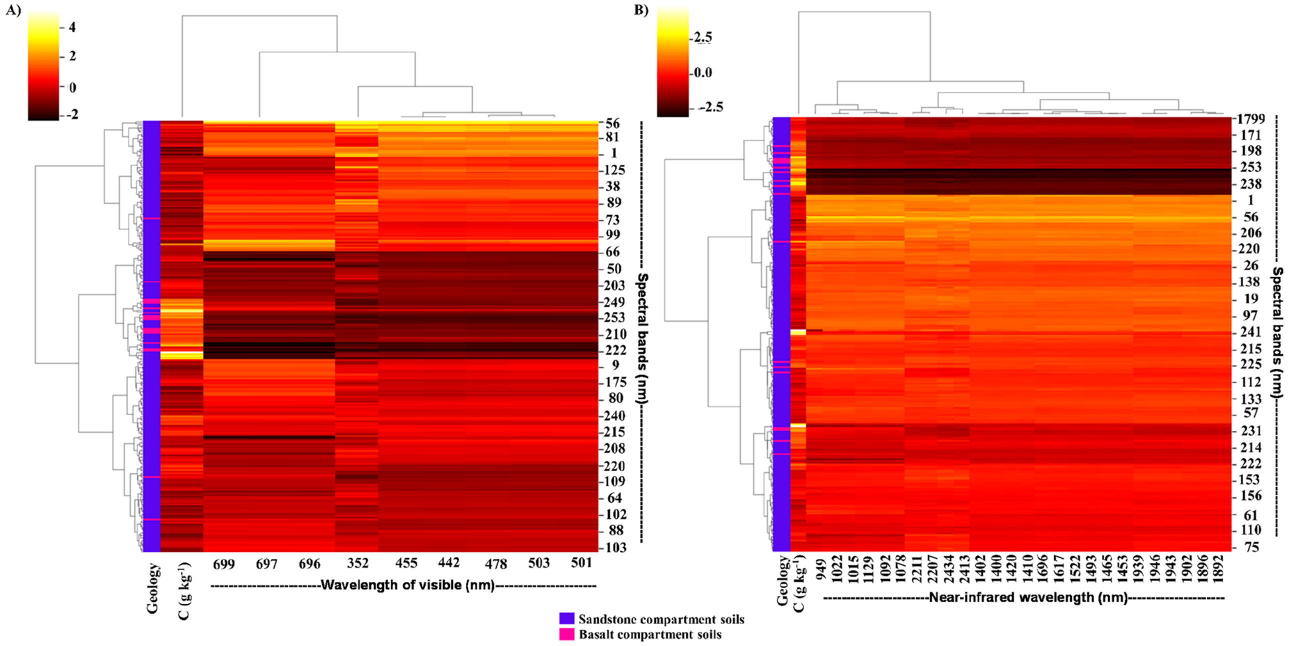

Figure 4. Cluster analysis with the heatmap for the most important spectral ranges for C estimation, as indicated by the stepwise modeling, C contents, and geological and geomorphological compartments for soils in the Western Plateau of São Paulo in the Vis (**A**) and NIR ranges (**B**). The color legend is a table that associates the colors used in the heatmap with the values of the main variable. The heatmaps were generated using the R language, gplots package, version 4.2.2 (https://www.R-project.org/).

reflectances are observed for soils originating from basalt, which have higher C means (Table 1). Spectral curves with higher reflectance are observed for sandstone soils (Fig. 3), which, in turn, have lower C means (Table 1).

Some samples of the basalt compartment were not grouped, being dispersed among the sandstone compartment samples. In theory, all samples from each compartment should form a single group. However, maps and scales for collecting the points were different. The points were chosen and collected based on previous experiences in smaller areas[40,41] to represent all the variability of 13 million hectares. Moreover, agronomic data present high variability in space and, therefore, present different behaviors from those expected.

Total C estimation and mapping

Estimation by random forest (RF)

Table 2 shows that the testing stage had lower metrics than the training stage. Heil et al.[83] also observed a reduction in metrics in the prediction stage for soil C and iron content. The C estimate for the Vis–NIR range showed the best metrics, with MAE of 2.50, RMSE of 3.40, and R^2 of 0.74 for training, and MAE of 3.81, RMSE of 4.88, and R^2 of 0.54 for testing. The metrics for the C estimate in the NIR range (MAE = 2.46, RMSE = 3.41, and R^2 = 0.73) were similar to those observed for the Vis–NIR range in the training stage. However, the R^2 in the NIR range was higher than that observed for the Vis–NIR range in the testing stage.

Morellos et al.[30] used the cubist ML algorithm and obtained an RMSE value of 2.18 and an R^2 of 0.78 for estimating organic C in a Luvisol in Germany. Gelsleichter et al.[19] studied the estimate C by RF and obtained an R^2 of 0.84 for soils in the Itatiaia National Park, Brazil. The results found in the literature are higher than those obtained in this study. However, the high variability of the area can lead to higher errors and more erratic predictions. However, they are still efficient when comparing the cost–benefit ratios [28,84,85]. Stenberg[13] pointed out that samples from sandy soils in the testing and/or calibration stages can result in worse attribute estimates. Sand has a small total surface compared to clay and OM, which can cause more erratic effects due to high absorption[12].

Figure 5 shows the regression between observed and estimated data. The estimate using the three ranges (Vis, NIR, and Vis–NIR) showed that values up to 10 g kg^{-1} of C, observed and estimated by RF, are very close to each other and the line. Values of C > 10 g kg^{-1} had more dispersion between the observed data and those estimated by RF. It may indicate that the algorithm has little sensitivity to learn and estimate higher C content. Therefore, the sample grid in areas with a higher C content must be densified to minimize errors. The wide differences between minimum and maximum values can make the prediction more erratic, as it is a large area, with a high geological and geomorphological variability.

In general, ML shows better results than traditional methods, but it can be considered of low interpretability[86]. The importance of the variables used in the estimation and the number of times they were mentioned in the rules must be evaluated to overcome this possible difficulty[87–89]. It enables the visualization of how the models were generated to interpret them (Fig. 5).

The C contents observed and estimated by RF for the Vis, NIR, and Vis–NIR ranges were subjected to geostatistical analysis for detailed mapping (Table 3). The C data observed and estimated by RF in the Vis range were adjusted to the exponential model. The C data estimated by RF in the NIR and Vis–NIR ranges adjusted to a spherical model.

Bahia et al.[18] worked with C mapping for a small area of the State of São Paulo and adjusted the spherical model to the data. La Scala Jr. et al.[26] and Bahia et al.[90] studied the soil CO_2 flux in areas of the State of São Paulo and also adjusted the spherical model to the data, pointing out that the data agreed with those observed in the literature. The spherical model is considered the most common for adjusting soil variables[56,91], as they present abrupt variations along the landscape[92].

The data estimated by RF presented shorter ranges and lower spatial dependence. The results showed that the algorithm has a high sensitivity to capture the heterogeneity of the data. The highest C contents were found in the south, southwest, and east edges, where basalt compartments are located (Fig. 6). The lowest C contents were found in the other WPSP regions, especially in the central region, where sandstones are located. The estimates using DRS and RF could detect these wide variations of C contents in the sandstone.

The data maps estimated by RF showed similar behavior to the observed C map (Fig. 6A–C). However, the algorithm tends to underestimate the C contents in both geological and geomorphological compartments. As indicated by the metrics, the estimation model that associated the Vis–NIR ranges generated the C map most similar to the observed map. The association of ranges allowed a better capture of transitions in the compartments, especially in the sandstone compartments. The C content estimation by Vis and NIR in the sandstone compartment was generalized although presenting similar maps to that observed.

		Training			Testing		
		MAE	RMSE	R^2	MAE	RMSE	R^2
C (g kg^{-1})	Vis	3.14	4.27	0.67	3.74	4.86	0.54
	NIR	2.46	3.41	0.73	3.69	4.73	0.59
	Vis–NIR	2.50	3.40	0.74	3.81	4.88	0.54

Table 2. Evaluation metrics of random forest models to estimate the total C (g kg^{-1}) for the Vis, NIR, and Vis–NIR ranges at a depth of 0.00–0.20 m for soils in the Western Plateau of São Paulo. *Vis* Visible, *NIR* near-infrared, *MAE* mean absolute error, *RMSE* root mean square error.

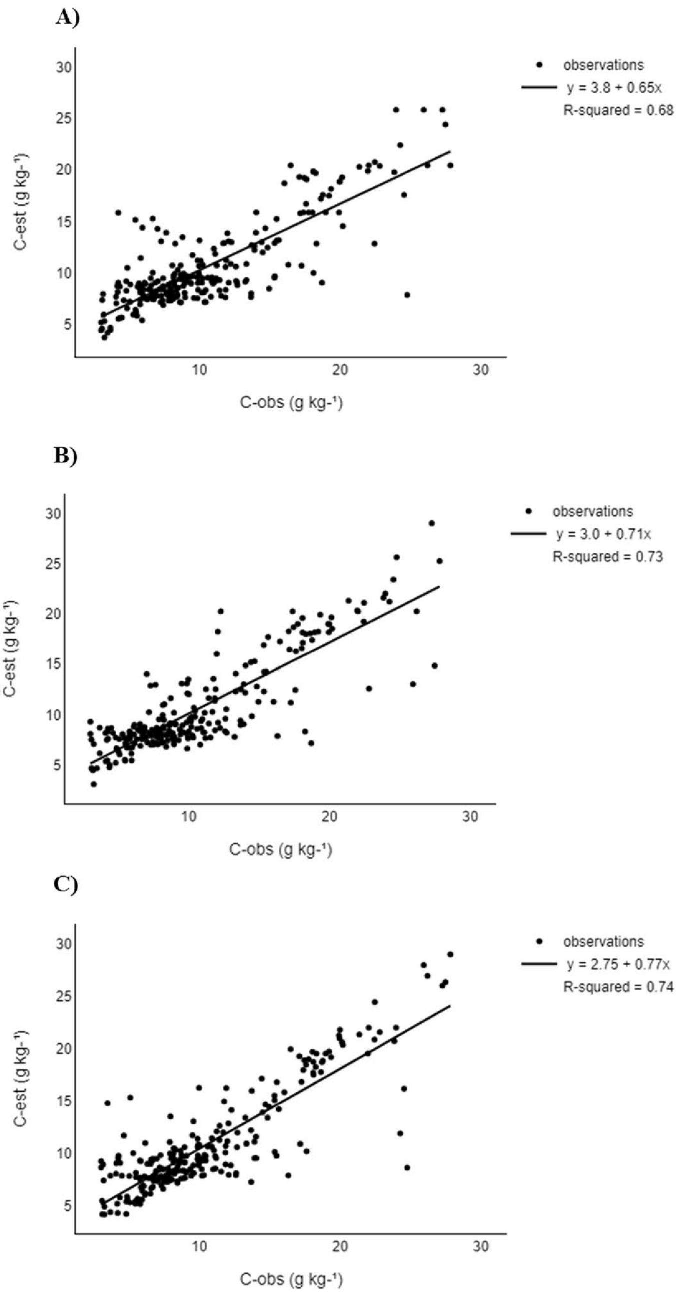

Figure 5. Regression analysis between predicted C (g kg^{-1}) data by the (**A**) Vis (visible), (**B**) NIR (near-infrared), and (**C**) Vis–NIR ranges by random forest and observed data for soils in the Western Plateau of São Paulo.

	Attribute	Model	C_0	$C_0 + C$	A_0 (m)	R^2	SQR	DSD
Observed	C (g kg^{-1})	Exp	19.4	58.22	5.95	0,82	5.32	0.66
Estimated RF	C (g kg^{-1})–Vis	Exp	3.68	7.44	0.25	0.78	2.24	0.50
	C (g kg^{-1})–NIR	Sph	4.93	9.87	0.93	0.90	2.65	0.50
	C (g kg^{-1})–Vis–NIR	Sph	5.88	13.62	1.01	0.91	3.50	0.56

Table 3. Geostatistical metrics of the total C contents observed and estimated by the Vis (visible), NIR (near-infrared), and Vis–NIR ranges by random forest (RF) at a depth of 0.00–0.20 m for soils of the Western Plateau of São Paulo. *Exp* exponential model, *Sph* spherical model, C_0 nugget effect, $C_0 + C$ sill, A_0 range, *SQR* sum of squared residuals, *DSD* degree of spatial dependence.

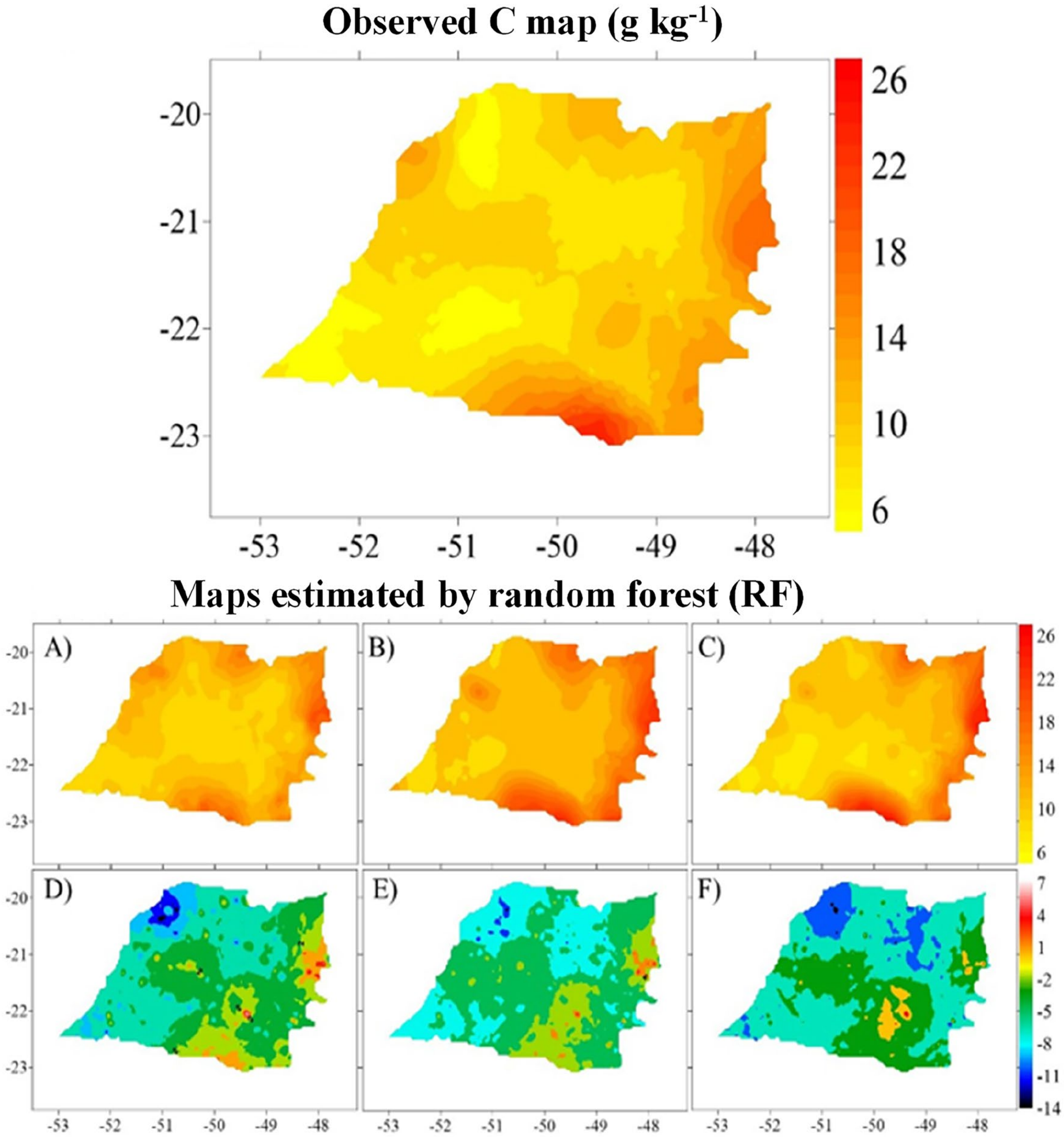

Figure 6. Observed C maps and estimated by random forest based on the ranges of the spectral curves Vis (**A**), NIR (**B**), and Vis–NIR (**C**) and their respective error maps (**D**–**F**) for soils of the Western Plateau of São Paulo. *Vis* visible, *NIR* near-infrared. The semivariograms and maps were generated using the R language, geostat package, version 4.2.2 (https://www.R-project.org/).

The superiority of mathematical modeling depends on the database. The choice of modeling depends on the database size and professional expertise. The RF analysis was carried out using the database reduced by stepwise analysis for practical and data processing purposes. The selection of specific wavelengths allows the researcher to work with a smaller database and sensors that capture wavelengths less frequently. Models derived from smaller spectral libraries can provide more accurate predictions of C content, surpassing predictions obtained with a model derived from a large spectral library[93].

The use of the most robust ML algorithms manages to capture the non-linearity of the data, besides being able to adapt to the database and improve the model accuracy[94]. ML techniques capture valuable spatial information that is not captured by environmental covariates, and the inclusion of this information improves the overall predictive performance[95]. These algorithms associated with DRS allow detailed mapping of C content in a precise and fast way. Indirect techniques, such as DRS, are efficient in capturing the spatial variability of C, being faster and more efficient, as well as less costly, than conventional techniques[18,85]. The maps generated using these estimates can assist in defining public policies for land use, reducing CO_2 emissions, and implementing C-credit policies.

Other potential technologies, such as UAV (unmanned aerial vehicles)-based NIR, also offer quality assessment and monitoring of soil C[96] and plant biomass[97]. Aldana-Jague et al.[96] investigated the potential of UAV multi-spectral imagery (480–1000 nm) for estimating the OC content in bare cultivated soils at a high spatial resolution (12 cm) and concluded that the methodology has a clear potential for use in precision agriculture or monitoring important soil properties following management changes. The authors obtained a coefficient of determination of 95% for the validation and an RMSE of 0.21% C based on cross-validation despite a sampling design that was not fully optimized for spectral calibration or spatial mapping. In this context, remote sensors attached to UAVs have been successfully calibrated using ground control points with known reflectance[98,99]. Cruciol et al.[100] investigated calibration approaches of UAV-based near-infrared digital images and observed that the conversion of visible and near-infrared digital images into reflectance promoted a decrease in the coefficient of variation of the spectral data for all visible bands.

Estimation by partial least squares regression (PLSR)

Table 4 shows the metrics obtained for the partial least squares regression (PLSR). The errors given by MAE, RMSE, and R^2 values are slightly higher in this prediction stage. Bahia et al.[18] estimated the total C by PLSR and obtained an R^2 of 0.82 for the Vis range and 0.88 for the Vis–NIR range in a small study area within WPSP. Asgari et al.[20] studied the estimate of organic C in Iran and obtained an R^2 of 0.74 for the Vis–NIR range. Viscarra Rossel et al.[81] obtained an R^2 of 0.86 for the NIR range and 0.91 for the Vis–NIR range. The extension and high variability of WPSP should be considered although the results are higher than those found in these studies.

The Vis range presented the best metrics for this modeling, with an R^2 of 0.64, MAE of 3.49, and RMSE of 4.40 for calibration, and R^2 of 0.58, MAE of 3.52, and RMSE of 6.65 for prediction (Table 4). The Vis range was followed by the Vis–NIR range, with an R^2 of 0.52, MAE of 3.40, and RMSE of 4.59 for calibration, and R^2 of 0.52, MAE of 3.48, and RMSE of 6.42 for prediction. The NIR range presented the worst metrics, mainly in the prediction stage, with an R^2 of 0.32, MAE of 3.52, and RMSE of 5.56 for calibration, and R^2 of 0.35, MAE of 3.59, and RMSE of 7.75 for prediction. Bahia et al.[18] explained that the metrics obtained for the Vis range can be justified by the good relationship between C and iron oxides. The relationships between these oxides and C contents were observed in this study and also by La Scala Jr. et al.[26] and Bahia et al.[15,90].

Unlike RF and regardless of the higher or lower C contents, the observed and estimated data showed higher dispersion along the line in the regression analysis (Fig. 7). The entire spectral curve was used for the PLSR modeling, as the calibration and prediction models need a wide variety of data. Wavelengths can have wide variations that influence the predictive capacity of the chemometric analysis[101]. Also, Wold et al.[101] explained that the PLSR analysis is based on data homogeneity, and very sudden variations of dilution in the curves or noise can lead to errors. In other words, working with only a few wavelengths or doing extreme preprocessing can favor errors in the estimation of attributes. Gelsleichter et al.[19] studied the preprocessing of spectral curves for C estimation and pointed out that it can decrease the predictive potential of PLSR.

The comparison between PLR with RF modeling showed that the latter algorithm had a better performance than the multivariate chemometric analysis. Morellos et al.[30] used RF and PLSR and observed that ML presents a better performance for estimating C. The authors stated that the efficiency of this model is due to its high capacity to deal with the nonlinear pattern of the data. Viscarra Rossel et al.[14] Keskin et al.[102] and Gelsleichter et al.[19] made the same considerations, corroborating the results observed in this study.

Regarding the lengths considered the most important for each range, the range between 420 to 480 nm is considered as important for C prediction. PLSR does not consider the length of 352 nm and above 500 nm as important for the estimation. In the NIR, RF considers wavelengths between 1000 to 2000 nm as important, while PLSR considers wavelengths above 2000 nm as more important. The ranges considered important for C prediction were wide compared to those found in the literature. Viscarra Rossel et al.[81] cited wavelengths of 400–2198 nm in the Vis–NIR range and 1100–2498 nm in the NIR range. Demattê et al.[103] cited several wavelengths, such as 825, 853, 1100, 1138, 1449, 1524, 1650, 1706, 1730, 1754, 1930, 1961, 2033, 2135, 2137, 2275, 2316, 2307, 2336, 2381, and 2469 nm. These authors corroborate with the results of this study regarding the specific wavelengths and the fact that more wavelengths in the NIR range may be important for estimating C.

The geostatistical analysis for detailed mapping (Table 5) showed that the C data estimated by PLSR were adjusted to three different models: exponential (Vis range), spherical (Vis–NIR range), and Gaussian (NIR range) (Table 5).

The closest ranges were obtained for the observed C data (5.95 m) and estimated by PLSR for the Vis (5.11 m), NIR (3.15 m), and Vis–NIR ranges (6.01 m), with the highest values of spatial dependence (DSD). The longest ranges for the estimated data indicate higher underestimations, generating a more homogeneous database.

The maps of data estimated by PLSR showed similar behavior to the map of observed C (Fig. 8A–C). The data estimated by the association of ranges (Vis–NIR) and NIR could not capture the variations between sandstone and basalt compartments, tending to overestimate C contents in basalt regions, which have more C, and

		Calibration			Prediction		
		MAE	RMSE	R^2	MAE	RMSE	R^2
C (g kg^{-1})	Vis	3.49	4.40	0.64	3.52	6.65	0.58
	NIR	3.52	5.56	0.32	3.59	7.75	0.35
	Vis–NIR	3.40	4.59	0.52	3.48	6.42	0.52

Table 4. Metrics for evaluating models by partial least squares regression to estimate the total C (g kg^{-1}) for the Vis, NIR, and Vis–NIR ranges at a depth of 0.00–0.20 m for soils of the Western Plateau of São Paulo. *Vis* Visible, *NIR* near-infrared, *MAE* mean absolute error, *RMSE* root mean square error.

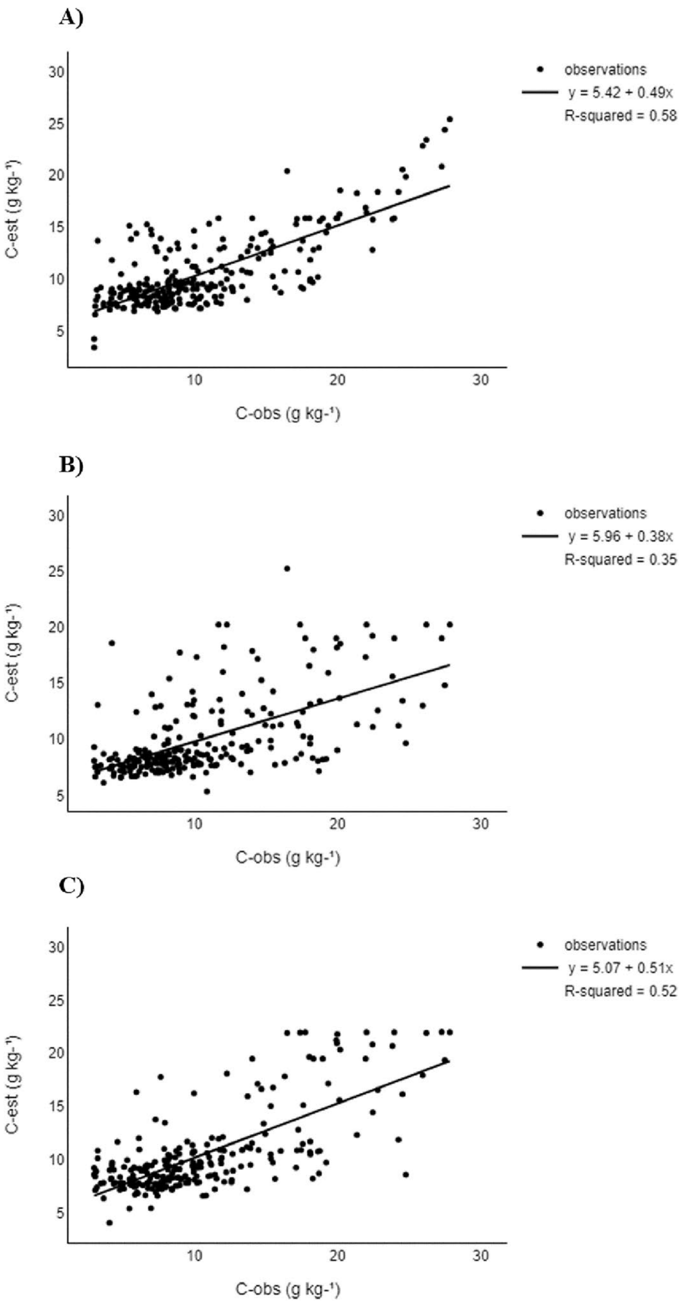

Figure 7. Regression analysis between the C (g kg^{-1}) data predicted by the (**A**) Vis (visible), (**B**) NIR (near-infrared), and (**C**) Vis–NIR ranges by partial least squares regression and observed data for soils in the Western Plateau of São Paulo.

	Attribute	Model	C_0	$C_0 + C$	A_0 (m)	R^2	SQR	DSD
Observed	C (g kg^{-1})	Exp	19.4	58.22	5.95	0.82	5.32	0.66
Estimated PLSR	C (g kg^{-1})–Vis	Exp	29.3	98.60	5.11	0.70	4.10	0.70
	C (g kg^{-1})–NIR	Gau	15.8	44.95	3.15	0.92	3.80	0.74
	C (g kg^{-1})–Vis–NIR	Sph	16.22	37.96	6.01	0.83	4.45	0.57

Table 5. Geostatistical metrics of the total C contents observed and estimated by the Vis (visible), NIR (near-infrared), and Vis–NIR ranges by partial least squares regression (PLSR) at a depth of 0.00–0.20 m for soils of the Western Plateau of São Paulo. *Exp* exponential model, *Sph* spherical model, *Gau* Gaussian model, C_0 nugget effect, $C_0 + C$ sill, A_0 range, *SQR* sum of squared residuals, *DSD* degree of spatial dependence.

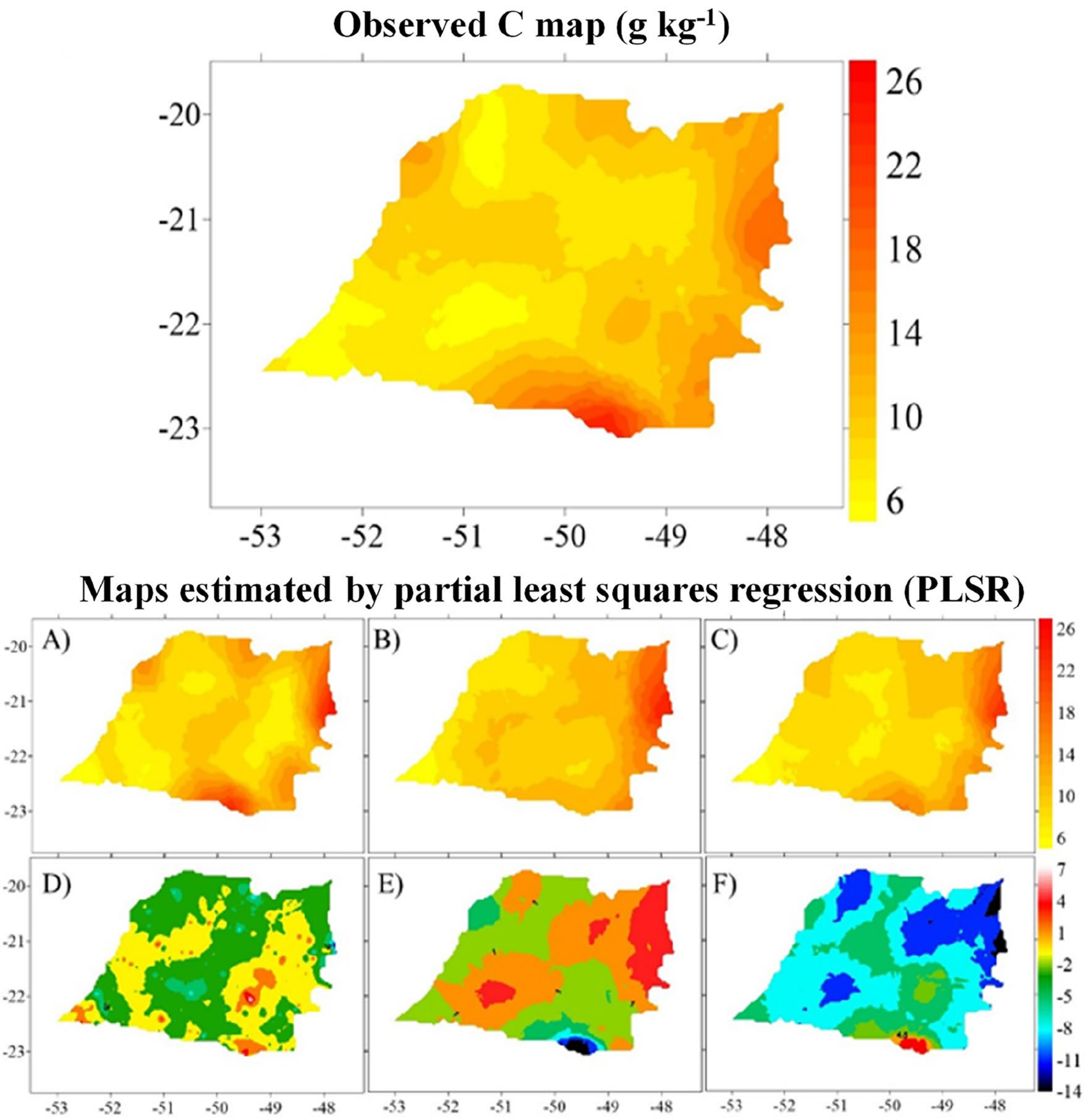

Figure 8. Observed C maps and estimated by partial least squares regression based on the ranges of the spectral curves Vis (**A**), NIR (**B**), and Vis–NIR (**C**) and their respective error maps (**D**–**F**) for soils of the Western Plateau of São Paulo. *Vis* visible, *NIR* near-infrared. The semivariograms and maps were generated using the R language, geostat package, version 4.2.2 (https://www.R-project.org/).

underestimate these contents in sandstone regions, with less C content. The southwest and south regions of WPSP have an extension of the basalt compartment with a higher C content (map of observed total C). However, these ranges in the PLSR analysis did not capture this variation.

Following the best PLSR metrics, the estimated data map for the Vis range showed the most similar patterns to the observed map (Fig. 8). The data in the basalt compartment followed the same patterns as the observed map. The data in the sandstone compartment were overestimated mainly in the central region of WPSP. Bahia et al.[18] mapped the total C content in a small area within WPSP and observed that PLSR underestimates the maximum C values, while it is more assertive for regions with lower C contents.

PLSR is a technique successfully used to predict numerous soil attributes, using spectral curves in various regions of the world. Using PLSR, Fernandes et al.[38] estimated soil mineral contents, Camargo et al.[61] estimated available and adsorbed P content, and Bahia et al.[18] estimated C and N contents. These studies, carried out in Brazilian soils, demonstrate the potential and diversity of the technique. Several other studies at an international level with the most varied soil attributes can also be mentioned[20,28,94].

Conclusions

The highest C contents are associated with more clayey soils, oxidic mineralogy, higher total pore volume, and lower soil density in highly dissected basalt compartments.

Soils with a higher reflectance have lower C content.

The most important wavelengths for estimating carbon are 352, 696, 697, 503, 501, 455, 442, and 478 nm in the visible range and 2211, 2413, 1696, 1617, 2207, 1946, 1465, 1400, 1078, 1022, 1902, 1696, 1493, 1410, 1943, 1939, 1896, 1892, 1522, 1453, 1402, 1420, 1129, 1092, and 949 nm in the mid-infrared range.

The random forest algorithm associated with the Vis–NIR spectral range is more efficient in estimating C content in tropical soils. The random forest algorithm associated with the Vis–NIR spectral range allowed the construction of a map of the estimated C content more similar to the observed map, pointing out that it can be used to define public policies for land use, reduce CO_2 emissions, and implement C credit policies in an economic and fast way.

Implications of the study: the importance of using a spectral sensor for pedogenesis in soil C mapping

This study provides an alternative strategy to the usual methodologies for evaluating and mapping soil characteristics, since traditional methods for quantifying soil properties involve field sampling and laboratory analysis, which are later mapped using interpolation to transform point data into a surface, making the process expensive, time-consuming, and laborious, and may not provide accurate information for large spatial areas[9,10]. Thus, spectral pedology using proximal sensors has been gaining prominence with the growing need for soil mapping with higher spatial resolution, presenting higher applicability in countries with large territorial extensions, such as Brazil[16]. It is a very promising technology that relates soil studies to the interpretation of electromagnetic spectra. This technology allows the discrimination of soil attributes[14,104], and the information is added to digital mapping. In addition, the adoption of new methodologies, such as diffuse reflectance spectroscopy (DRS), and robust mathematical techniques, such as machine learning and geostatistics, work together to make research of this magnitude possible[16,105].

Soil mineralogy and texture are directly related to C stock due to the physical properties of the clay surface[106,107]. Inadequate and unplanned land use, related to a lack of knowledge of the dynamics of C entry and exit from the soil, increases CO_2 emissions by 31%[32]. In this context, the topic addressed by this research is based on environmental sustainability and food security. According to England and Viscarra-Rossel[105], C stocks in the soil can be stabilized or increased through the identification and implementation of agronomic and environmental management. However, soil C shows high variability, requiring quantification to verify changes in its stock due to human activities, which demand a large sample volume[93]. Therefore, the study of C concentrations and their dynamics in the landscape through mapping could serve as a basis for countless future decisions, identifying, for example, where the soil has the highest potential to store C naturally, thus guiding actions for the mitigation of CO_2 emissions from agricultural food production systems[108].

The use of geology and geomorphology to compartmentalize the study area helps to determine the spatial variability of soil attributes. In this context, we understand that knowledge of these attributes in an area of 13 million hectares with economic relevance and heterogeneous in numerous characteristics can guide not only work in the State of São Paulo but also in neighboring states that have similar conditions, demonstrating the future potential of this research.

Therefore, we present here unprecedented results when considering methodology, territorial extension, and mapping. Thus, associated research in other Brazilian regions, using different methodologies and mathematical techniques, or showing how these maps can be used through field experimentation, are extremely important for validating the results. All the generated data is now part of a regional database and can be used as a basis for other research, guiding a pioneering development of research and innovation and contributing to the economical and sustainable advancement of science in Brazil.

Data availability

The data that support the findings of this study are available from the corresponding author, [Moitinho, M.R], on reasonable request.

Received: 31 August 2023; Accepted: 1 July 2024
Published online: 27 July 2024

References

1. Lal, R. Carbon sequestration. *Trans. R. Soc.* **1492**, B363815–B363830 (2008).
2. Le Quéré, C. *et al.* Global carbon budget 2018. *Earth Syst. Sci.* **10**, 2141–2194 (2018).
3. Batjes, N. H. Harmonized soil property values for broad-scale modelling (WISE30sec) with estimates of global soil carbon stocks. *Geoderma* **269**, 61–68 (2016).
4. Paustian, K. *et al.* Climate-smart soils. *Nature* **532**, 49–57 (2016).
5. Smith, P. *et al.* Land-management options for greenhouse gas removal and their impacts on ecosystem services and the sustainable development goals. *Annu. Rev. Environ. Resour.* **44**, 255–286 (2019).
6. Lal, R. Food security impacts of the "4 per Thousand" initiative. *Geoderma* **374**, 114427 (2020).
7. Dasgupta, S. & Robinson, E. J. Z. Attributing changes in food insecurity to a changing climate. *Sci. Rep.* **12**, 4709 (2022).
8. Minasny, B. *et al.* Soil carbon 4 per mille. *Geoderma* **292**, 59–86 (2017).
9. LECO Corporation, St Joseph, MI, USA. Carbon, Hydrogen, and Nitrogen in Coal. https://eu.leco.com/images/Analytical-Application-Library/CHN628_COAL_203-821-403.pdf (2013).
10. Sato, J. H. *et al.* Methods of soil organic carbon determination in Brazilian savannah soils. *Scientia Agricola* **71**, 302–308 (2014).
11. Intergovernmental Panel on Climate Change—IPCC. *Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge University Press, 2022).
12. Clark, R. N. Spectroscopy of rocks and minerals and principles of spectroscopy. Manual of Remote Sensing, in: Remote Sensing for the Earth Sciences: *Man. Remote Sens.* 3–58 (1999).

13. Stenberg, B., Viscarra-Rossel, R. A., Mouazen, A. M. & Wetterlind, J. Visible and near infrared spectroscopy in soil science. *Soil Sci.* 163–215 (2010).
14. Viscarra Rossel, R. A., Bui, E. N., Caritat, P. & Mckenzie, N. J. Mapping iron oxides and the color of Australian soil using visible—near—infrared reflectance spectra. *J. Geophys. Res.* **115**, 1–13 (2010).
15. Bahia, A. S. R. S., Marques, J. Jr. & Siqueira, D. S. Procedures using diffuse reflectance spectroscopy for estimating hematite and goethite in Oxisols of São Paulo, Brazil. *Geoderma Reg.* **5**, 150–156 (2015).
16. 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., ten 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., Sá Barretto Sampaio, E. V., 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, É. F. M., Grunwald, S., Vasques, G. M., Marques Jr, J., Silva, A. J., Barreto, M. C. V., Nóbrega, G. N., Silva, M. Z., Souza, S. F., Valladares, G. S., Viana, J. H. M., Silva Terra, F., 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. D. L. M. S., Ruivo, M. D. 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. The brazilian soil spectral library (BSSL): A general view, application and challenges. *Geoderma* **354**, 113793 (2019).
17. Bellon-Maurel, V. & McBratney, A. Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils—Critical review and research perspectives. *Soil Biol. Biochem.* **43**, 1398–1410 (2011).
18. Bahia, A. S. R. S., Marques, J. Jr., La Scala, N., Cerri, C. E. P. & Camargo, L. A. Prediction and mapping of soil attributes using diffuse reflectance spectroscopy and magnetic susceptibility. *Soil Sci. Soc. Am. J.* **81**, 1450–1462 (2017).
19. Gelsleichter, Y. A. *et al.* Machine learning algorithms for soil properties prediction with treated vis–NIR spectrums from the Itatiaia National Park. *Preprints* **1**, 1–20 (2019).
20. Asgari, N., Ayoubi, S., Demattê, J. A. M. & Dotto, A. C. Carbonates and organic matter in soils characterized by reflected energy from 350–25000 nm wavelength. *J. Mt. Sci.* **17**, 1636–1651 (2020).
21. Somarathna, P. D. S. N., Minasny, B., Malone, B. P., Stockmann, U. & McBratney, A. B. Accounting for the measurement error of spectroscopically inferred soil carbon data for improved precision of spatial predictions. *Sci. Total Environ.* **631–632**, 377–389 (2018).
22. Barbieri, D. M., Marques, J. Jr., Alleoni, L. R. F., Garbuio, F. J. & Camargo, L. A. Hillslope curvature, clay mineralogy, and phosphorus adsorption in an Alfisol cultivated with sugarcane. *Sci. Agric.* **66**, 819–826 (2009).
23. Montanari, R., Marques, J. Jr., Campos, M. C. C., Souza, Z. M. & Camargo, L. A. Caracterização mineralógica de Latossolos em diferentes feições do relevo na região de Jaboticabal, SP. *Revista Ciência Agronômica* **41**, 191–199 (2010).
24. Carmo, D. A. B. *et al.* Cor do solo na identificação de áreas com diferentes potenciais produtivos e qualidade de café. *Pesquisa Agropecuária Brasileira* **51**, 1261–1271 (2016).
25. Rasmussen, C. *et al.* Beyond clay: Towards an improved set of variables for predicting soil organic matter content. *Biogeochemistry* **137**, 297–306 (2018).
26. La Scala Jr, N., Marques, J. Jr., Pereira, G. T. & Cora, J. E. Short-term temporal changes in the spatial variability model of CO_2 emissions from a Brazilian bare soil. *Soil Biol. Biochem.* **32**, 1459–1462 (2000).
27. Padarian, J., Minasny, B. & McBratney, A. B. B. Using Google's cloud-based platform for digital soil mapping. *Comput. Geosci.* **83**, 80–88 (2015).
28. Viscarra-Rossel, R. A. Parles: Software for chemometric analysis of spectroscopic data. *Chemometr. Intell. Lab. Syst.* **90**, 72–83 (2008).
29. Breiman, L. Random forest. *Mach. Learn.* **45**, 5–32 (2001).
30. Morellos, A. *et al.* Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. *Biosyst. Eng.* **152**, 104–116 (2016).
31. Quinton, J. N., Govers, G., Van Oost, K. & Bardgett, R. D. The impact of agricultural soil erosion on biogeochemical cycling. *Nat. Geosci.* **3**, 311–314 (2010).
32. Lybbert, T. J. & Sumner, D. A. Agricultural technologies for climate change in developing countries: Policy options for innovation and technology diffusion. *Food Policy* **37**, 114–123 (2012).
33. 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* **8**, 6 (2013).
34. Barbieri, D. M. *Atributos físicos, químicos e mineralógicos de um Latossolo Vermelho eutroférrico sob dois sistemas de colheita de cana-de-açúcar*. Tese (Doutor em Ciência do Solo) UNESP, Jaboticabal, SP (2011).
35. Fernandes, L. A. Mapa litoestratigráfico da parte oriental da Bacia Bauru (PR, SP, MG), escala 1:1.000.000. Bol. Paraná. *Geocience* 53–66 (2004).
36. Fernandes, L. A., Castro, A. B. & Basilici, G. Seismites in continental sand sea deposits of the Late Cretaceous Caiuá Desert, Bauru Basin, Brasil. *Sediment. Geol.* **199**, 61–64 (2007).
37. Silva, L. S. *et al.* Spatial variability of iron oxides in soils from Brazilian sandstone and basalt. *Catena* **185**, 104258 (2020).
38. Fernandes, K., Marques, J. Jr., Bahia, A. S. R. S. , Demattê, J. A. M. & Ribon, A. A. Landscape-scale spatial variability of kaolinite-gibbsite ratio in tropical soils detected by diffuse reflectance spectroscopy. *Catena* **195**, 104795 (2020).
39. Dinkins, D. P. & Jonessoil, C. Sampling strategies. *Agriculture and Natural Resources*. MT200803AG New 4/08 (2008).
40. Teixeira, D. D. B. *et al.* Spatial variability of soil CO_2 emission in a sugarcane area characterized by secondary information. *Sci. Agric.* **70**, 195–203 (2013).
41. Marques, J. Jr., Alleoni, L. R. F., Teixeira, D. D. B., Siqueira, D. S. & Pereira, G. T. Sampling planning of micronutrients and aluminium of the soils of. *Geoderma Reg.* **4**, 91–99 (2015).
42. Teixeira, P. C., Donagemma, G. K., Fontana, A. & Teixeira, W. G. *Manual de métodos de análise de solo* 573p (Embrapa, 2017).
43. Raij, B. V. *Análise química para avaliação da fertilidade de solos tropicais*. Campinas, Instituto Agronômico (2001).
44. Mehra, O. P. & Jackson, M. L. Iron oxide removal from soils and clays by a dithionite-citrate system buffered with sodium bicarbonate. *Clays Clay Miner.* **7**, 317–327 (1958).
45. Camargo, O. A., Moniz, A. C., Jorge, J. A. & Valadares, L. M. A. S. *Métodos de análise química, mineralógica e física dos solos do Instituto Agronômico de Campinas*. Instituto Agronômico, Campinas (1986).
46. Schwertmann, U. Relations between iron oxides, soil color, and soil formation. In *Soil Color* (eds Bigham, J. M. & Ciolkosz, E. J.) 51–69 (Soil Science Society of America, 1993).
47. Jackson, M. L. *Soil Chemical Analysis*, 2nd edn. 930 (1985).
48. Mehra, O. P. & Jackson, M. L. Iron oxide removed from soils and clays by dithionitecitrate system buffered with sodium bicarbonate. *Clays Clay Miner.* **7**, 1317–1327 (1960).
49. Norrish, K. & Taylor, R. M. The isomorphous replacement of iron by aluminium in soil goethites. *J. Soil Sci.* **12**, 294–305 (1961).
50. Kämpf, N. & Schwertmann, U. Relações entre óxidos de ferro e cor em solos cauliníticos do Rio Grande do Sul. *Revista Brasileira de Ciência do Solo* **7**, 27–31 (1982).
51. Breiman, L. ST4_Method_Random_Forest. *Mach. Learn.* **45**, 5–32 (2001).
52. Yan, F., Shangguan, W., Zhang, J. & Hu, B. Depth-to-bedrock map of China at a spatial resolution of 100 meters. *Sci. Data* **7**, 1–13 (2020).

53. Chagas, C. S., Carvalho Jr, W., Bhering, S. B. & Calderano Filho, B. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. *Catena* **139**, 232–240 (2016).
54. Pahlavan-Rad, M. R. *et al.* Prediction of soil water infiltration using multiple linear regression and random forest in a dry flood plain, eastern Iran. *Catena* **194**, 104715 (2020).
55. Pedregosa, F. G. V. *et al.* Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
56. Vieira, S. R. Geoestatística em estudos de variabilidade espacial do solo. In *Tópicos em Ciência do Solo* (eds. Novais, R. F., Alvarez, V. V. H. & Schaefer, C. E.) 1–54 (Sociedade Brasileira de Ciência do Solo, 2000).
57. Warrick, A. W. & Nielsen, D. R. Spatial variability of soil physical properties in the field. In *Applications of Soil Physics* (ed. Hillel, D.) 319–344 (Academic, 1980).
58. Sanchez, R. B., Marques, J. Jr., Souza, Z. M., Pereira, G. T. & Filho, M. V. M. Variabilidade espacial de atributos do solo e de fatores de erosão em diferentes pedoformas. *Bragantia* **68**, 1095–1103 (2009).
59. Cunha, P., Marques, J. Jr., Curi, N., Pereira, G. T. & Lepsch, I. F. Superfícies geomórficas e atributos de Latossolos em um sequência arenítico-basáltica da região de Jaboticabal (SP). *Revista Brasileira de Ciência do Solo* **29**, 81–90 (2005).
60. Siqueira, D. S. *et al.* Correlation of properties of Brazilian Haplustalfs with magnetic susceptibility measurements. *Soil Use Manag.* **26**, 425–431 (2010).
61. Camargo, L. A., Marques, J. Jr., Pereira, G. T. & Bahia, A. S. R. S. Clay mineralogy and magnetic susceptibility of Oxisols in geomorphic surfaces. *Sci. Agric.* **71**, 244–256 (2014).
62. Curi, N. & Franzmeier, D. P. Effect of parent rocks on chemical and mineralogical properties of some Oxisols in Brazil. *Soil Sci. Soc. Am. J.* **51**, 153–158 (1987).
63. Mendes, T. *Cenários de uso e manejo para conservação do estoque de carbono do solo no Estado do Maranhão*. Tese (Doutorado em Ciência do Solo)—Faculdade de Ciências Agrárias e Veterinárias, Universidade Estadual Paulista, Jaboticabal (2015).
64. Leal, F. T. *et al.* Characterization of potential CO_2 emissions in agricultural areas using magnetic susceptibility. *Sci. Agric.* **72**, 535–539 (2015).
65. Parron, L. M., Garcia, J. R., Oliveira, E. B., Brown, G. G. & Prado, R. B. Estoques de carbono no solo como indicador de serviços ambientais. In *Serviços Ambientais Em Sistemas Agrícolas e Florestais Do Bioma Mata Atlântica* 92–100 (Embrapa, 2015).
66. Assad, E. D. *et al.* Changes in soil carbon stocks in Brazil due to land use: Paired site comparisons and a regional pasture soil survey. *Biogeosciences* **10**, 6141–6160 (2013).
67. Saiz, G. *et al.* Variation in soil carbon stocks and their determinants across a precipitation gradient in West Africa. *Glob. Change Biol.* **18**, 1670–1683 (2012).
68. Vasconcelos, V., Souza, D., Martins, E., Carvalho Junior, O. A., Marques Jr, J., Silva, D. S., Couto Junior, A. F., Guimarães, R. F., Gomes, R. A. T. & Reatto, A. Modelo de evolução pedogeomorfológica da Serra da Canastra, MG. *Revista Brasileira de Geomorfologia* **14**, 197–212 (2013).
69. Schaefer, C. E. G. R., Fabris, J. D. & Ker, J. C. Minerals in the clay fraction of Brazilian Latosols (Oxisols): A review. *Clay Minerals* **43**, 137–154 (2008).
70. Kämpf, N. O ferro no solo. In *reunião sobre ferro em solos inundados, 1, Goiânia, 1988. Anais... Goiânia, Embrapa - CNPAF*, 35–71 (1988).
71. Genú, A. M., Demattê, J. A. M. & Nanni, M. R. Caracterização e comparação do comportamento espectral de atributos do solo obtidos por sensores orbitais (ASTER e TM) e terrestre (IRIS). *Ambiência Guarapuava* **9**, 279–288 (2013).
72. Viscarra Rossel, R. A. & Bouma, J. Soil sensing: A new paradigm for agriculture. *Agric. Syst.* **148**, 71–74 (2016).
73. Genú, A. M., Demattê, J. A. M. & Fiorio, P. R. Análise espectral de solos da Região de Mogi-Guaçú (SP). *Semina* **31**, 1235–1244 (2010).
74. Demattê, J. A. M. *et al.* The Brazilian Soil Spectral Library (BSSL): A general view, application and challenges. *Geoderma* **354**, 113793 (2019).
75. Fang, Q. *et al.* Visible and near-infrared reflectance spectroscopy for investigating soil mineralogy: A review. *J. Spectrosc.* **2018**, 1–14 (2018).
76. Zhao, D., Zhao, X., Khongnawang, T., Arshad, M. & Triantafilis, J. A Vis-NIR spectral library to predict clay in australian cotton growing soil. *Soil Sci. Soc. Am. J.* **82**, 1347 (2018).
77. Chicati, M. S., Nanni, M. R., Chicati, M. L., Furlanetto, R. H., Cezar, E. & Oliveira, R. B. Hyperspectral remote detection as an alternative to correlate data of soil constituents. *Remote Sens. Appl. Soc. Environ.* **16** (2019).
78. Dufréchou, G., Grandjean, G. & Bourguignon, A. Geometrical analysis of laboratory soil spectra in the short-wave infrared domain: Clay composition and estimation of the swelling potential. *Geoderma* **243–244**, 92–107 (2015).
79. Bishop, J. L., Lane, M. D. & Dyar, M. D. J. B. A. Reflectance and emission spectroscopy study of four groups of phyllosilicates: Smectites, kaolinite-serpentines, chlorites and micas. *Clay Minerals* **43**, 35–54 (2008).
80. Silvero, N. E. Q. *et al.* Effects of water, organic matter, and iron forms in mid-IR spectra of soils: Assessments from laboratory to satellite-simulated data. *Geoderma* **375**, 114480 (2020).
81. Viscarra Rossel, R. A., Walvoort, D. J. J., McBratney, A. B., Janik, L. J. & Skjemstad, J. O. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma* **131**, 59–75 (2006).
82. Demattê, J. A. M., Epiphanio, J. C. & Formaggio, A. R. Influência da matéria orgânica e de formas de ferro na reflectância de solos tropicais. *Bragantia* **62**, 451–464 (2003).
83. Demattê, J. A. M. *et al.* Genesis and properties of wetland soils by VIS-NIR-SWIR as a technique for environmental monitoring. *J. Environ. Manag.* **197**, 50–62 (2017).
84. Gómez-Robledo, L. *et al.* Using the mobile phone as munsell soil-colour sensor: An experiment under controlled illumination conditions. *Comput. Electron. Agric.* **99**, 200–208 (2013).
85. Aitkenhead, M. *et al.* Digital RGB photography and visible-range spectroscopy for soil composition analysis. *Geoderma* **313**, 265–275 (2018).
86. Padarian, J., Minasny, B. & McBratney, A. B. Machine learning and soil sciences: A review aided by machine learning tools. *Soil* **6**, 35–52 (2020).
87. Martin, M. *et al.* Evaluation of modelling approaches for predicting the spatial distribution of soil organic carbon stocks at the national scale. *Geoderma* **223**, 97–107 (2014).
88. Schillaci, C. *et al.* Modelling the topsoil carbon stock of agricul- tural lands with the Stochastic Gradient Treeboost in a semi-arid Mediterranean region. *Geoderma* **286**, 35–45 (2017).
89. Khanal, S., Fulton, J., Klopfenstein, A., Douridas, N. & Shearer, S. Integration of high resolution remotely sensed data and machine learning techniques for spatial prediction of soil properties and corn yield. *Comput. Electron. Agric.* **153**, 213–225 (2018).
90. Bahia, A. S. R. S. *et al.* Iron oxides as proxies for characterizing anisotropy in soil CO_2 emission in sugarcane areas under green harvest. *Agric. Ecosyst. Environ.* **192**, 152–162 (2014).
91. Cambardella, C. A. *et al.* Field-scale variability of soil properties in Central Iowa. *Soil Sci. Soc. Am. J.* **58**, 1501–1508 (1994).
92. Issaks, E. H. & Srivastava, R. M. *A Introduction to Applied Geostatistics*, 561 (University Press, 1989).
93. Guerrero, C. *et al.* Do we really need large spectral libraries for local scale SOC assessment with NIR spectroscopy?. *Soil Tillage Res.* **155**, 501–509 (2016).
94. Marsland, S. W. *Machine Learning An Algorithmic Perspective*. 2nd ed. (2015).

95. Beguin, J., Fuglstad, G., Mansuy, N. & Paré, D. Predicting soil properties in the Canadian boreal forest with limited data: Comparison of spatial and non-spatial statistical approaches. *Geoderma* **306**, 195–205 (2017).
96. Aldana-Jague, E., Heckrath, G., Macdonald, A., van Wesemael, B. & Van Oost, K. UAS-based soil carbon mapping using VIS-NIR (480–1000 nm) multi-spectral imaging: Potential and limitations. *Geoderma* **75**, 55–66 (2016).
97. Jenal, A. *et al.* Investigating the potential of a newly developed UAV-based VNIR/SWIR imaging system for forage mass monitoring. *PFG* **88**, 493–507 (2020).
98. Berni, J. A. J., Zarco-Tejada, P. L., Suárez, L. & Fereres, E. Thermal and narrowband multiespectral remote sensing for vegetation monitoring from unmanned aerial vehicle. *IEEE Trans. Geosci. Remote Sens.* **47**(2009), 722–738 (2009).
99. Crusiol, L. G. T. *et al.* Semi professional digital camera calibration techniques for Vis/NIR spectral data acquisition from an unmanned aerial vehicle. *Int. J. Remote Sens.* **38**, 2717–2736 (2017).
100. Crusiol, L. G. T., Nanni, M. R., Furlanetto, R. H., Cezar, E. & Silva, G. F. C. Reflectance calibration of UAV-based visible and near-infrared digital images acquired under variant altitude and illumination conditions. *Remote Sens. Appl. Soc. Environ.* **18**, 100312 (2020).
101. Wold, S., Sjöström, M. & Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemometr. Intell. Lab. Syst.* **58**, 109–130 (2001).
102. Keskin, H., Grunwald, S. & Harris, W. G. Digital mapping of soil carbon fractions with machine learning. *Geoderma* **339**, 40–58 (2019).
103. Demattê, J. A. M., Morgan, C. L. S., Chabrillat, S., Rizzo, R., Franceschini, M. H. D., Terra, F. S., Vasques, G. M. & Wetterlind, J. Spestral sensing from ground to space in soil science: stat of the art, applications, potencial and perspectives. In *Remote Sensing Handbook: Land Resources Monitoring, Modeling and Mapping with Remote Sensing* 661–732 (2016).
104. Arrouays, D., Poggio, L., Guerrero, O. A. S. & Mulder, V. L. Digital soil mapping and GlobalSoilMap: Main advances and ways forward. *Geoderma Reg.* **21**, e00265 (2020).
105. England, J. R. & Viscarra-Rossel, R. A. Proximal sensing for soil carbon accounting. *Soil* **4**, 101–122 (2018).
106. Churchman, G. J., Singh, M., Scalpel, A., Sarkar, B. & Bolan, N. Clay minerals as the key to the sequestration of carbon in soils. *Clays Clay Miner.* **68**, 135–143 (2020).
107. Prout, J. M., Shepherd, K. D., McGrath, S. P., Kirk, G. J. & Haefele, S. M. What is a good level of soil organic matter? An index based on organic carbon to clay ratio. *Eur. J. Soil Sci.* **72**, 2493–2503 (2021).
108. Padarian, J., Minasny, B., McBratney, A. & Smith, P. Soil carbon sequestration potential in global croplands. *PeerJ* **10**, e13740 (2022).

Acknowledgements

To the São Paulo Research Foundation (FAPESP, Brazil) for providing funding for this study (Processes No. 2015/20692-0 and 2017/05477-0). To the Coordination for the Improvement of Higher Education Personnel (CAPES, Brazil) (Process No. 149940). To the Pro-Rector of Research of São Paulo State University (PROPE/UNESP) (Call No. 15/2014). To the National Council for Scientific and Technological Development (CNPq, Brazil) (Process No. 402796/2016-0).

Author contributions

K.F., J.M.J., A.A.R., G.M.A., M.R.M., D.L.D.D., A.S.R.S.B., D.M.S.O. Contributions Conceptualization: K.F.; J.M.J.; A.A.R.; G.M.A.; M.R.M.; D.L.D.D.; D.M.S.O. Methodology: K.F.; J.M.J.; A.A.R.; G.M.A.; M.R.M.; D.L.D.D.; D.M.S.O. Software: K.F.; G.M.A. Validation formal analysis: K.F.; J.M.J.; A.A.R.; G.M.A. Investigation: K.F.; J.M.J.; A.A.R.; G.M.A.; M.R.M.; D.L.D.D.; D.M.S.O. Resources: J.M.J. Data curation: K.F.; J.M.J.; G.M.A. Writing—original draft preparation: K.F.; J.M.J.; A.A.R.; G.M.A.; M.R.M.; D.L.D.D.; D.M.S.O. Writing—review: K.F.; J.M.J.; A.A.R.; G.M.A.; M.R.M.; D.L.D.D.; D.M.S.O. Supervision: J.M.J. Project administration: J.M.J. Design of the work: K.F.; J.M.J.; A.A.R.; G.M.A.; M.R.M.; D.L.D.D. Revising it critically: K.F.; J.M.J.; A.A.R.; G.M.A.; M.R.M.; D.L.D.D.; D.M.S.O. All authors have read and agreed to the last version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.R.M.