

Article

Strategies for the Development of Spectral Models for Soil Organic Matter Estimation

Everson Cezar ¹, Marcos Rafael Nanni ¹, Luís Guilherme Teixeira Crusiol ^{1,2}, Liang Sun ^{2,*},
Mônica Sacioto Chicati ¹, Renato Herrig Furlanetto ¹, Marlon Rodrigues ¹, Rubson Natal Ribeiro Sibaldelli ³,
Guilherme Fernando Capristo Silva ¹, Karym Mayara de Oliveira ¹ and José A. M. Demattê ⁴

¹ Department of Agronomy, State University of Maringá, Maringá, PR 87020-900, Brazil; ecezar2@uem.br (E.C.); mnrnanni@uem.br (M.R.N.); pg53403@uem.br (L.G.T.C.); monicasacioto@gmail.com (M.S.C.); pg53830@uem.br (R.H.F.); pg53382@uem.br (M.R.); guilhermecapristo51@gmail.com (G.F.C.S.); pg54640@uem.br (K.M.d.O.)

² Key Laboratory of Agricultural Remote Sensing, Ministry of Agriculture/CAAS-CIAT Joint Laboratory in Advanced Technologies for Sustainable Agriculture—Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences, Beijing 100081, China

³ Mathematician, Statistical Specialist, Londrina, PR 86001-970, Brazil; rubson.sibaldelli@embrapa.br

⁴ Department of Soil Science, University of São Paulo, Piracicaba, SP 13418-900, Brazil; jamdemat@usp.br

* Correspondence: sunliang@caas.cn



Citation: Cezar, E.; Nanni, M.R.; Crusiol, L.G.T.; Sun, L.; Chicati, M.S.; Furlanetto, R.H.; Rodrigues, M.; Sibaldelli, R.N.R.; Silva, G.F.C.; Oliveira, K.M.d.; et al. Strategies for the Development of Spectral Models for Soil Organic Matter Estimation. *Remote Sens.* **2021**, *13*, 1376. <https://doi.org/10.3390/rs13071376>

Academic Editor: Asa Gholizadeh

Received: 5 March 2021

Accepted: 31 March 2021

Published: 2 April 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: Visible (V), Near Infrared (NIR) and Short Waves Infrared (SWIR) spectroscopy has been indicated as a promising tool in soil studies, especially in the last decade. However, in order to apply this method, it is necessary to develop prediction models with the capacity to capture the intrinsic differences between agricultural areas and incorporate them in the modeling process. High quality estimates are generally obtained when these models are applied to soil samples displaying characteristics similar to the samples used in their construction. However, low quality predictions are noted when applied to samples from new areas presenting different characteristics. One way to solve this problem is by recalibrating the models using selected samples from the area of interest. Based on this premise, the aim of this study was to use the spiking technique and spiking associated with hybridization to expand prediction models and estimate organic matter content in a target area undergoing different uses and management. A total of 425 soil samples were used for the generation of the state model, as well as 200 samples from a target area to select the subsets (10 samples) used for model recalibration. The spectral readings of the samples were obtained in the laboratory using the ASD FieldSpec 3 Jr. Sensor from 350 to 2500 nm. The spectral curves of the samples were then associated to the soil attributes by means of a partial least squares regression (PLSR). The state model obtained better results when recalibrated with samples selected through a cluster analysis. The use of hybrid spectral curves did not generate significant improvements, presenting estimates, in most cases, lower than the state model applied without recalibration. The use of the isolated spiking technique was more effective in comparison with the spiked and hybridized state models, reaching r^2 , square root of mean prediction error (RMSEP) and ratio of performance to deviation (RPD) values of 0.43, 4.4 g dm^{−3}, and 1.36, respectively.

Keywords: PLSR; hybrid curves; reflectance; recalibration; spiking

1. Introduction

Soil is a heterogeneous system, displaying complex processes and mechanisms that are difficult to comprehensively understand. Many conventional analytical techniques have been employed in an attempt to establish a direct relationship between soil and soil properties [1]. Knowledge concerning the soil system, its interactions and quality has been systematically supported through routine analyses which, although reliable, involve the collection of large numbers of samples, as well as laborious analysis processes. A high

economic cost is also noted, due to the use of chemical reagents, laboratory equipment, and personnel, with the added generation of hazardous waste [2].

The search for agile, low cost, accurate, and low environmental impact analytical tools is key in soil studies, mainly those on a large scale. The search for the development of cleaner and more economically viable technologies that allow for analytical repeatability as well as a minimum of sample preparation [3,4], has become one of the great demands of the 21st century. Most likely due to these reasons, Visible, Near Infrared and Short Wave Infrared (VNIR-SWIR) spectroscopy has been considered a possible alternative to improve or replace conventional laboratory soil analysis methods [5,6]. This technique meets all the desirable characteristics and leads to less onerous soil attribute predictions [7].

VNIR-SWIR spectra are signals containing information concerning chemical, physical, and mineralogical soil characteristics. Models constructed through multivariate statistical regression techniques [8], termed chemometric analyses, can be applied in order to take advantage of this information.

Despite the existence of this sophisticated method and the fact that VNIR-SWIR spectroscopy presents a robust physicochemical basis, the samples used to construct a model must be capable of representing both the chemical and physical attributes of the area to which it will be applied. In the present study, we attempted to calibrate models from spectral libraries containing a high number of samples, which should be representative and encompass the entire expected variability for the study area [9]. However, according to Viscarra Rossel et al. [10] and Wetterlind et al. [11], calibrations obtained from large sample sizes do not always guarantee direct applicability to new environments.

In this case, model application in non-sampled locations and, therefore, without any representatives in the spectral library, can lead to errors during the prediction phase, resulting in low accuracy [12]. Current studies have demonstrated that recalibration of so-called regional, state, or global models with samples from an area to be assessed (spiking) is the best way to overcome this problem [13–15].

For this, some of the soil samples from the target area are scanned by means of spectroscopy and are also analyzed in the laboratory using the classic reference method of analysis to estimate chemical, physical, and mineralogical attributes. Subsequently, these samples are added to the original calibration matrix which represents different scales and thus the prediction model is recalibrated. The domain of the new model starts to be expanded and it contains the variability of the target area. This process can often improve the model accuracy for the remaining soil samples from target area [16,17]. The key issue in this process is that, although it is known that the spiking technique presents promising potential for model improvements and soil attribute estimations, it does not always lead to good results. Studies such as those performed by Guerrero et al. [15] and Guy et al. [18] have shown that the achievement of satisfactory results is mainly linked to the characteristics of the target area. Estimates usually present low efficiency when the samples are spectrally very different from each other, as well as very different from those used to generate the regional, state, or global models.

In this case, the use of the so-called hybrid spectra seems to be the best way to fill the spectral space between the model and target area samples. The hybrid spectrum (average spectrum) is an “artificial” spectrum constructed using a sample from the spectral library and a local sample [19]. It is important to highlight that as these hybrid spectra share characteristics of both sets of samples, they could increase the number of spectra in the spectral space located between both sets of samples without additional analytical cost.

With this strategy, it is expected that this spectral space, filled with hybrid samples, will have greater relevance within the recalibrated model, allowing a better estimate for the remaining samples of the target area. Studies such as the one by Guerrero and Zornoza [19] seem to confirm these assumptions, offering positive results, but which should be addressed with a larger set of data in relation to the work done by the researchers.

In this context, the aim of the present study was to assess the effects of the spiking technique and hybridization in the recalibration of a state model and their reflexes on

organic matter estimations in areas that are under different uses and management. We chose to work with organic matter, due to the fact that it is an important indicator of the physical and chemical quality of the soil [20,21]. On the other hand, procedures such as Walkley and Black [22], traditionally used to quantify this attribute uses dangerous chemical reagents such a chromium solution [23].

Associated with this, there is a strong demand in the State of Paraná for chemical analyzes of determination of organic matter, since monitoring the content of this element is important for the development of management practices that will improve and maintain the productivity of agricultural soils [24].

Facing the aforementioned, this study is expected to contribute to understanding how we can expand the use of VNIR-SWIR spectroscopy in the estimation of organic matter in new areas, based on a minimum number of soil samples and spectral data of this environment.

2. Materials and Methods

2.1. Soil Sampling

A total of 425 soil samples were collected from different areas submitted to different uses and management in the state of Paraná, Brazil (0–20 cm depth). The sampling points were selected based on the pedological and geological maps of the state of Paraná represented on a scale of 1:600.000 and 1:650.000, respectively [25,26]. The sites were chosen in such a way that there was no collection of repeated soil samples on similar source materials. This care was taken so that there was no influence from repeated samples in the spectral models to be adjusted. Use of soil as well as the management practices adopted in agricultural areas were recorded as the field work was being carried out. Subsequently, this information was inserted into a geographic information system for the creation of a georeferenced database and a spectral library of the state of Paraná. The predominant soil classes in the areas are *lixisols*, *cambisols*, *gleysols*, *ferralsols*, *arenosols*, *nitisols*, and *histosols* [27].

In addition, a total of 200 soil samples (0–20 cm depth) were collected in a specific area located in the Lobato municipality, northwestern of Paraná-Brazil, comprising 2500 ha. This area is basically occupied with remnants of forest and sugar cane crops. The soil classes from the area include *ferralsols*, *nitisols*, *lixisols*, *cambisols*, and *arenosols* [27]. The location of the target area within the state of Paraná is shown below (Figure 1).

2.2. Organic Matter and Spectral Analyses

The soil samples were oven dried at 45 °C for 24 h and subsequently sieved through a 2 mm mesh to be submitted to chemical analysis. Total organic carbon was determined following the Walkley and Black methodology [22]. The organic matter content was obtained by multiplying the total organic carbon by 1.724, since it is admitted that in the humus medium composition, carbon participates with 58% [28].

The organic matter attribute was chosen for spectral modelling because it is an important indicator of soil quality in Brazil and its traditional determination uses reagents which, without the correct final destination, may contaminate the environment.

For the determination of spectral readings, in addition to the drying process aforementioned, the samples were milled to homogenize soil particle size and reduce roughness effects [29]. The samples were then arranged in a 9 cm diameter and 1.5 cm high Petri dish and spectral readings were carried out using an ASD FieldSpec 3 JR spectroradiometer, which covers the spectral range from 350 to 2500 nm. The equipment was programmed to perform 50 readings per sample, generating an average spectral curve.

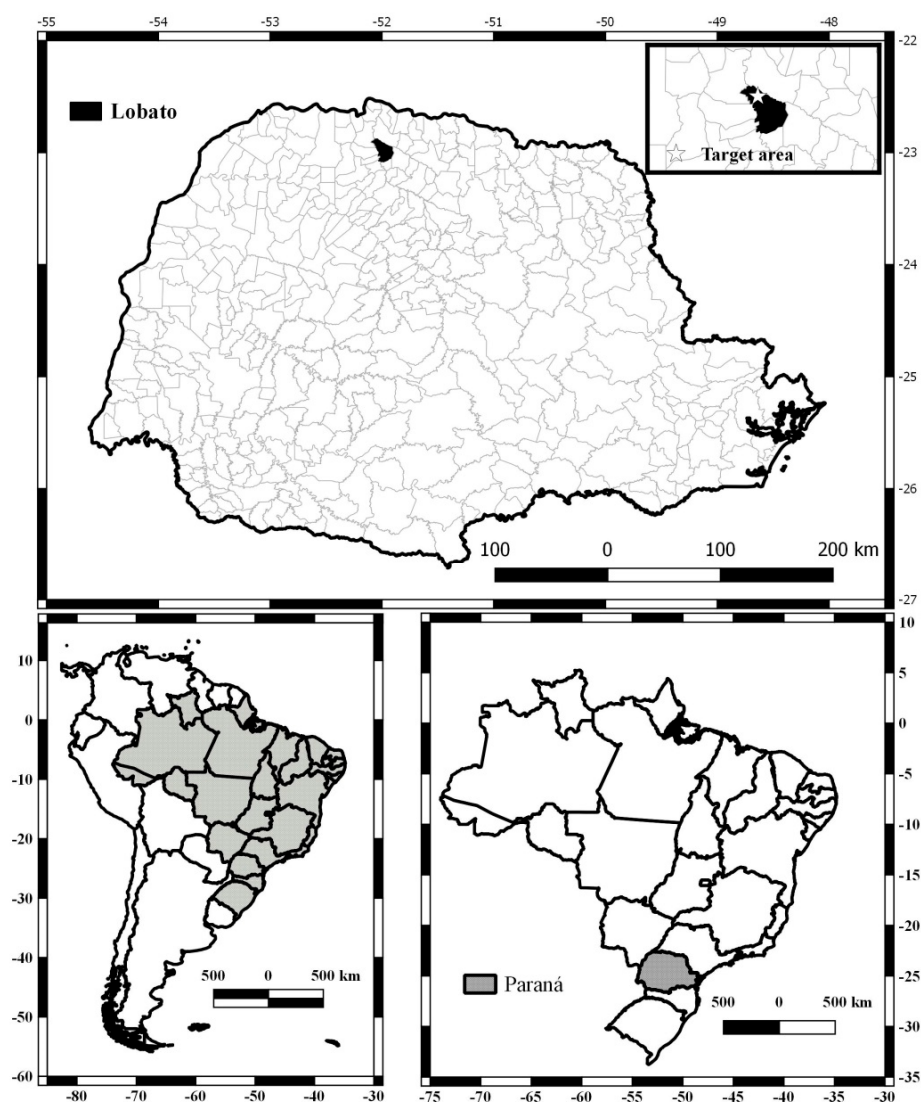


Figure 1. Location of the state of Paraná and the target area inserted in the northwestern region.

A standard white plate with 100% calibrated reflectance was used for data acquisition, according to the Labsphere Reflectance Calibration Laboratory [30]. The fiber optic reader was placed 8 cm upright from the sample support platform. The reading area was of approximately 2 cm². The light source used was a 650 W lamp with an uncollimated beam for the target plane, positioned 35 cm from the platform and at a 30° angle to the horizontal plane [31].

The spectral readings were repeated three times, with successive displacement of the Petri dish 120° clockwise, allowing for a full sample scan. Subsequently, the simple arithmetic means of the three readings was determined for each sample, as recommended by Nanni and Demattê [32].

2.3. Data Processing and Statistical Analyses

Raw spectral data were preprocessed to improve the stability of the regression models, as described by Lee et al. [33]. Each spectral curve was subjected to baseline correction and light scattering by the multiplicative scatter correction (MSC) method, according to Buddenbaum and Steffens [34]. For noise reduction, the Savitzky-Golay Smoothing method [35] was used, with the first derivative employed using seven smoothing points. The calibration models were constructed applying the partial least squares regressions (PLSR), using the Unscrambler version 10.3 software package (CAMO, Inc., Oslo, Norway).

The prediction performance of the models was assessed using the coefficient of determination (r^2), square root of mean prediction error (RMSEP), standard error (SEP), systematic error (BIAS) and ratio of performance to deviation (RPD), as described by D'Acqui et al. [13].

2.4. State Model

The state model (unspiked state model) was generated from a total of 425 soil samples collected in the state of Paraná. Its effectiveness in the estimation of organic matter was tested on 200 samples collected in the target area.

Recalibrated State Model

For this step, the spiking technique was used to recalibrate the state model with selected samples from the target area. Sample selection was performed based on spectral sample characteristics. The criterion of choice was based on the distribution of spectra from the set of samples from the target area within the spectral domain they occupy. In this way, we tried to use spectra that were at the limit of the spectral domain, as well as those that were in the center or even randomly distributed, with the objective of covering the entire spectral space.

The selection based on cluster analysis sought to group smaller samples with spectral similarity into smaller subsets. As it is an unsupervised analysis, a biased selection is discarded, in addition, recalibration with samples from different clusters may indicate that the best soil samples from a set should be employed in the Spiking technique. Initially recalibrations were tested with five and 10 samples, however, five samples proved to be an insufficient number (not presented).

A large number of samples was not tested in the recalibration, since such a procedure is not recommended, since routine analyzes would be required to determine organic matter, which would increase costs, contrary to the application of the spectroscopy technique. The selection criteria are presented below.

A total of 10 samples located at the periphery of the spectral space, comprising the first two main components (subset one), 10 samples located in the center of the spectral space, comprising the first two main components (subset two), 10 samples located along the spectral space consisting of the first two main components (subset three) and 10 samples belonging to different clusters (k-means) (subset four) were chosen, according to Cezar et al. [36].

In a second step, the state model was recalibrated with hybrid spectra, obtained using the dataset of the target area and of the state of Paraná. In order to obtain these spectra, the four selection criteria mentioned above were applied to both datasets. After the subset's selection, the simple means between the corresponding spectra was calculated for each criterion, obtaining 10 hybrid spectra for each criterion, with a total of 40 hybrid spectra. A general state model recalibration scheme is presented in Figure 2.

After state model recalibration, the model was used on an unknown data matrix (95% of the remaining samples of the target area), for performance and predictive ability testing.

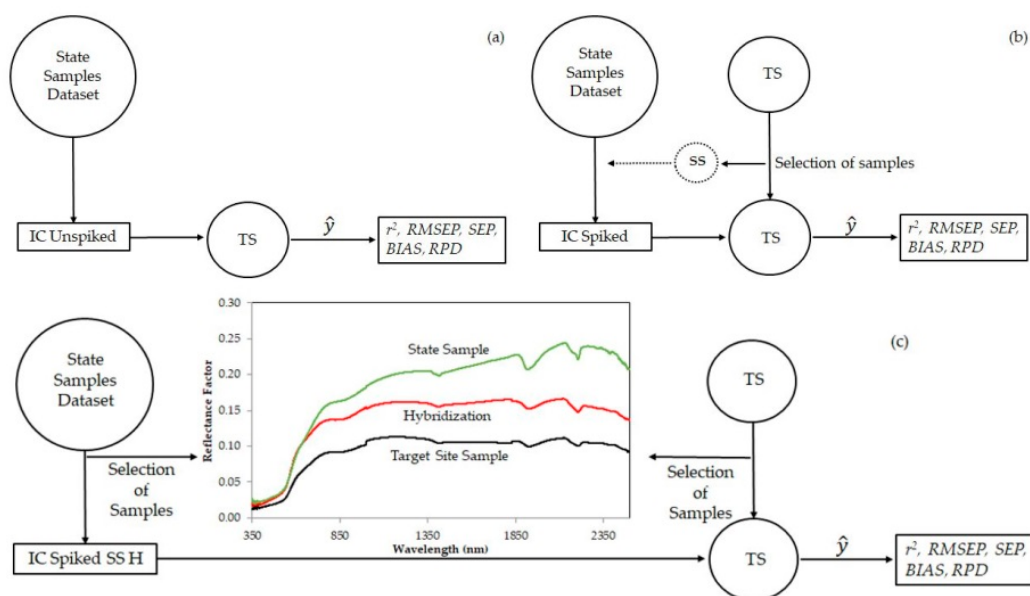


Figure 2. Scheme used to represent the experiment. (a) Unspiked initial calibration (IC) model constructed only with state samples; (b) Initial calibration spiked with a spiking subset (SS) selected from the target site (TS); (c) Initial calibration spiked with a spiking subset and hybridization (SSH) selected from the target site (TS) and state of Paraná. Adapted from Guerrero et al. [15].

3. Results and Discussions

3.1. Statistical Soil Sample Characterization

The results obtained from the descriptive analysis indicate that organic matter content is variable. Considering the entire state, the oscillation between the minimum and maximum values is high, reaching values above 60 g dm^{-3} (Figure 3). Compared to the set of samples from the target area, the standard deviation of the set of samples collected in the state of Paraná is higher, showing a smaller homogeneity among the data.

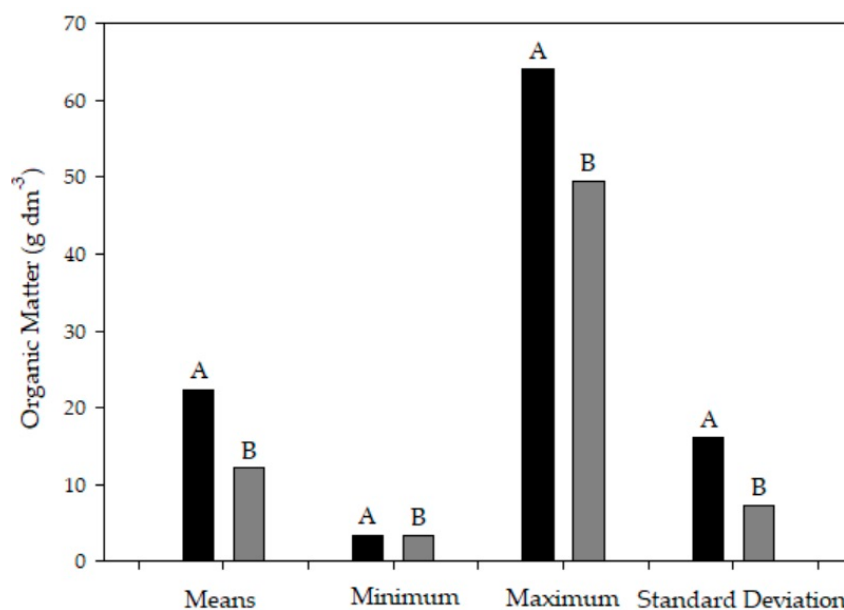


Figure 3. A: Sample set collected in state of Paraná ($n = 425$); B: Sample set collected in target area ($n = 200$); n = number of samples.

The high variability (around 72.23%, not presented) of the state of the Paraná samples is mainly due to the existence of differences in the state climatic conditions [37], leading to significant variations in organic matter content in several regions [36].

In addition, the managements applied in the agricultural areas also result in greater variability, with higher or lower organic matter accumulation on the soil surface. Therefore, those variations were expected, since both no-tillage and tillage planting are observed in the state of Paraná, with higher organic matter accumulation over the years in no-tillage planting, agreeing with Martínez et al. [38].

On the other hand, the sample set from the target area presented lower variability, around 59.23% (not presented). In this case, some factors such as climate and management were less relevant, due to the smaller size of the area, 2500 ha, mostly used for sugarcane plantations, which undergoes the same management during the crop cycle. At a lesser extent, some forest remnants are also observed.

3.2. Spectral Soil Sample Characterization

The spectral curves representative of the samples used for the model generation also showed inter-state differences, as well as differences between the state and the target area samples (Figure 4).

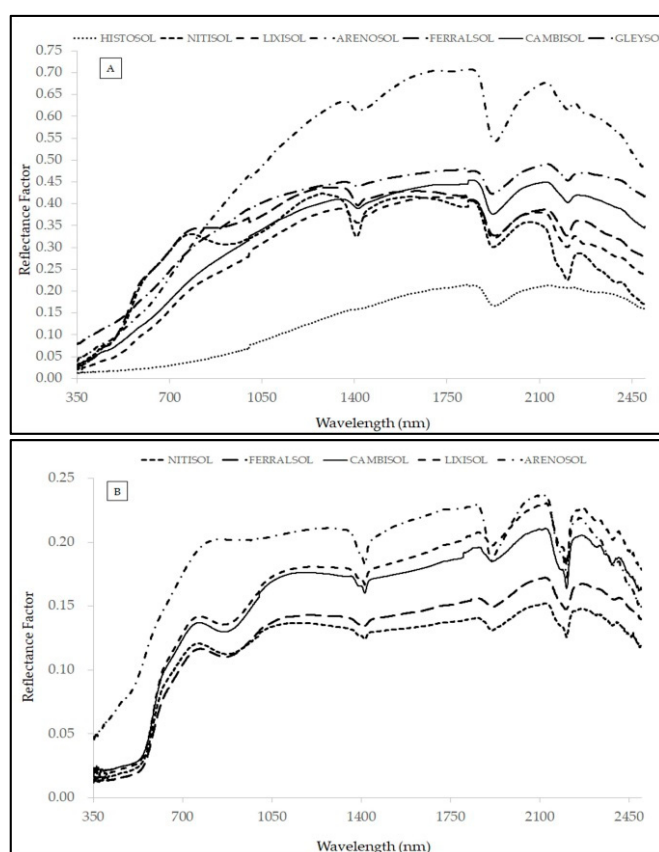


Figure 4. Mean spectral curves obtained for the sample set from the state of Paraná (A) and from the target area (B), separated by soil classes.

The average spectral curves for the samples collected in the target area are better defined, with distinct inter-sample spectral differentiation, mainly in wavelengths greater than 700 nm. On the other hand, the spectral curves for the Paraná samples are less differentiated (except for histosols and arenosols that have very different spectral behavior) and are better separated in wavelengths greater than 1900 nm. The reflectance factor of the

target area samples presents a lower amplitude, ranging from 0.02 to 0.25, while for the Paraná samples it ranges from 0.01 to 0.70.

This difference occurs mainly as a function of soil variability in Paraná State, as described in Section 2.1. In addition, soil use and management can lead to changes in spectral behavior, mainly due to variations in organic matter content, which can absorb electromagnetic radiation at all wavelengths, masking absorption bands generated by other elements [39]. This behavior can be observed through the evaluation of the spectral curve of samples collected in the remaining forest area, classified as histosols (Figure 4), which do not display absorption bands except in the 1900 nm region, characteristic for the presence of water [40].

Iron occurrence was the same for all target area samples (absorption around 900 nm), except for arenosols. This agrees with one of the materials of origin that form this soil, which displays low iron concentrations. On the other hand, the spectral responses of the Paraná samples indicated the presence of iron only for ferralsols and nitisols, which usually present this element above 150 g kg^{-1} of soil.

The other classes, besides presenting lower levels than the aforementioned soil, displayed impaired iron absorption bands due to the presence of organic matter, [41,42]. This was higher than 2% for 175 samples, with the potential to influence any spectral curve, as discussed by Baumgardner et al. [43] and Bilgili et al. [44].

3.3. Unspiked State Model

Calibration and Prediction

Table 1 displays the results obtained during model calibration and validation.

Table 1. Statistical parameters obtained during the calibration and validation phase of the unspiked state model.

Calibration							
<i>n</i>	r^2 ⁽¹⁾	RMSEC ⁽²⁾	SEC ⁽³⁾	BIAS ⁽⁴⁾	R ⁽⁵⁾	RPD ⁽⁶⁾	N° factors
425	0.86	10.2	10.2	6.05×10^{-4}	0.77	1.63	10
Prediction							
<i>n</i>	r^2	RMSEP ⁽⁷⁾	SEP ⁽⁸⁾	BIAS	R	RPD	N° factors
200	0.30	5.2 5.2	5.2	−0.43	0.67	1.42	7

⁽¹⁾ Determination coefficient; ⁽²⁾ Root-Mean-Square Error calibration; ⁽³⁾ Standard Error calibration; ⁽⁴⁾ Calibration Systematic Error; ⁽⁵⁾ Correlation Coefficient; ⁽⁶⁾ Ratio of Performance to Deviation; ⁽⁷⁾ Root-Mean-Square Error for prediction; ⁽⁸⁾ Standard error prediction; *n*: number of Paraná and target site soil samples.

Although the BIAS, correlation coefficient, and coefficient of determination results were satisfactory, the RPD value was below the ideal for use in agriculture, as discussed by Viscarra Rossel et al. [1], considered as presenting average precision. According to these researchers, the ideal values for use with agriculture would be above 3, where values from 2 to 3 are considered good, 1.5 to 2 average, and below 1.5, unsatisfactory. This finding is corroborated when using this model to estimate organic matter content in the target area, where a low ability to accurately predict these values is observed. The RPD in this case is noteworthy, at 1.42, relatively superior to the values obtained by Cezar et al. [36] in a similar study.

Therefore, it is evident that, even in the case of a large state model composed of 425 soil samples, the entire variability of the target area could not be determined, with no accurate representation of the organic matter contents present in the target area, agreeing with what was described by Viscarra Rossel et al. [45] and Guerrero et al. [15]. The presence of variability can be corroborated by the means of the average spectral curves for the target area, which present differences in terms of absorption bands, as well as oscillations in reflectance values for the different soil classes (Figure 4).

3.4. Spiked State Model/Spiked and Hybridized State Model

3.4.1. Recalibration

The statistical parameters presented a small improvement over the unspiked state model after recalibration of the state model through the spiking technique and spiking associated with hybridization (Table 2).

Table 2. Statistical parameters obtained by recalibration with samples from the target area ($n = 435$).

Spiked State Model							
Subset	r^2 ⁽¹⁾	RMSEC ⁽²⁾	SEC ⁽³⁾	BIAS ⁽⁴⁾	R ⁽⁵⁾	RPD ⁽⁶⁾	N° factors
1	0.87	9.8	9.8	6.91×10^{-4}	0.80	1.68	14
2	0.87	9.7	9.7	2.18×10^{-4}	0.79	1.66	14
3	0.86	9.9	9.9	-3.25×10^{-4}	0.78	1.65	14
4	0.88	9.6	9.6	-3.58×10^{-5}	0.80	1.72	14
Spiked and Hybridized State Model							
Subset	r^2 ⁽¹⁾	RMSEC ⁽²⁾	SEC ⁽³⁾	BIAS ⁽⁴⁾	R ⁽⁵⁾	RPD ⁽⁶⁾	N° factors
1	0.87	9.9	9.9	2.40×10^{-3}	0.79	1.66	14
2	0.88	9.6	9.7	9.46×10^{-5}	0.80	1.68	14
3	0.87	9.8	9.8	4.75×10^{-4}	0.78	1.66	11
4	0.87	9.7	9.7	2.41×10^{-4}	0.79	1.65	11

⁽¹⁾ Recalibration Determination Coefficient; ⁽²⁾ Recalibration Root-Mean-Square Error; ⁽³⁾ Recalibration Standard Error; ⁽⁴⁾ Recalibration Systematic Error; ⁽⁵⁾ Correlation Coefficient; ⁽⁶⁾ Residual Predictive Deviation; n : number of samples used in the recalibrated model.

When assessing the spiked state model, it is noted that the RMSEC is lower, ranging from 9.6 to 9.9 g dm⁻³, while the correlation coefficient reaches a maximum value of 0.80 for the models recalibrated with subset one and subset four. The RPD values are also relatively better, reaching a maximum of 1.72. These results indicate that the recalibration of the state models with some target area samples can lead to improvements in the statistical parameters, agreeing with Guerrero et al. [46] and Hong et al. [17].

Similar behavior was observed after the use of the spiked state model associated with hybridization for recalibration of the state models. However, no significant improvements were observed after the use of this technique. RMSEC values reached a maximum of 9.9 g dm⁻³, while RPD reached 1.66. It should be noted that, in both cases, BIAS was low, demonstrating a lack of bias for the generated models.

3.4.2. Prediction

The statistical parameters obtained by the recalibrated model during the prediction phase are presented in Table 3.

A relative improvement in organic matter content estimates is noted after the use of the spiked state model in a new area, in agreement with the one described by Brown et al. [9], Sankey et al. [12], and Wetterlind and Stenberg [14]. The RMSEP values were lower than those obtained for the unspiked state model, while the correlation coefficient and determination values were higher.

When compared to the work of Lazzaretti et al. [47], which estimated soil organic matter through NIR spectroscopy associated with an unspiked model in the southern region of Brazil, the results were also slightly higher. Emphasis should be given to the correlation coefficient, which ranged from 0.68 to 0.76 (Table 3), depending on the strategy used for selecting samples used in the recalibration of the state model, against 0.62 of the aforementioned work.

Table 3. Statistical parameters obtained during the prediction phase, using 95% of the remaining samples from the target area ($n = 190$).

Spiked State Model							
Subset	r^2 ⁽¹⁾	RMSEP ⁽²⁾	SEP ⁽³⁾	BIAS ⁽⁴⁾	R ⁽⁵⁾	RPD ⁽⁶⁾	N° factors
1	0.41	4.8	4.8	0.33	0.68	1.52	7
2	0.38	4.9	4.9	−0.41	0.70	1.28	7
3	0.37	4.9	5.3	1.20	0.68	1.37	7
4	0.43	4.4	4.7	0.94	0.76	1.36	6
Spiked and Hybridized State Model							
Subset	r^2 ⁽¹⁾	RMSEP ⁽²⁾	SEP ⁽³⁾	BIAS ⁽⁴⁾	R ⁽⁵⁾	RPD ⁽⁶⁾	N° factors
1	0.32	6.2	6.7	1.37	0.62	1.22	10
2	0.34	5.1	5.6	1.39	0.69	1.21	7
3	0.34	5.0	5.6	1.47	0.71	1.23	6
4	0.41	4.6	4.7	0.49	0.71	1.55	7

⁽¹⁾ Prediction Determination Coefficient; ⁽²⁾ Prediction Root-Mean-Square Error; ⁽³⁾ Prediction Standard Error; ⁽⁴⁾ Prediction Systematic Error; ⁽⁵⁾ Correlation Coefficient; ⁽⁶⁾ Residual Predictive Deviation; n : number of soil samples used in the prediction.

On the other hand, the results were inferior to those obtained by Lazaar et al. [48] who, working in Eastern Morocco with organic matter estimation by means of VNIR-SWIR spectroscopy, obtained r^2 equal to 0.93 and RMSE equal to 0.13. It should be noted that in this case, the prediction models as well as their validations were developed with a smaller number of samples (lower variability) when compared to the work carried out in Paraná State.

Similarly, the results were lower than those achieved by Qiao et al. [49], who developed organic matter prediction models for Chinese soils using hyperspectral data. These researchers obtained values of r^2 and RPD (prediction) equal to 0.61 and 5.53, whereas we obtained maximum values for these indices equal to 0.43 and 1.36, respectively, considering the spiked state model.

As already explained for Lazaaret al. [48], the number of samples used in the calibration (165) and validation (15) of this work is small when compared to that used for the study of Paraná soils. While we obtained a mean value for organic matter above 20 g dm^{−3} and a maximum value above 60 g dm^{−3} (Figure 3), the study by Qiao et al. [49] obtained a mean value of 2.60 g dm^{−3} and a maximum of 4.33 g dm^{−3}. These results demonstrate the difference between the organic matter content of tropical soils and other parts of the world.

Regarding the use of hybridization, this technique did not allow for improvements in organic matter content estimates when compared to the spiked state model, since lower quality indices were found in most cases (subset one, subset two, and subset three). Likewise, when comparing the recalibrated models using hybrid curves with the unspiked state model, only subset four presented better results, with RMSEP, BIAS, and the correlation coefficient equal to 4.6 g dm^{−3}, 1.55, and 0.71, respectively.

The lack of effectiveness of the hybridization to improve organic matter predictions lies in the fact that, although the spectral curves are different in terms of absorption and reflectance, as displayed in Figure 4, both datasets (from Paraná and the target area) are within the same spectral domain, demonstrating that these samples are not spectrally distant, agreeing with Nawar and Mouazen [50] and Hong et al. [17].

This assertion can be corroborated by means of the similarity map formed by principal component analysis scores. When the PCA model generated with the Paraná soil sample spectra was applied to the target area sample spectra, the scores were calculated and projected to the local site within the spectral space occupied by the state samples (Figure 5), similar to that obtained by Wetterlind and Stenberg [14].

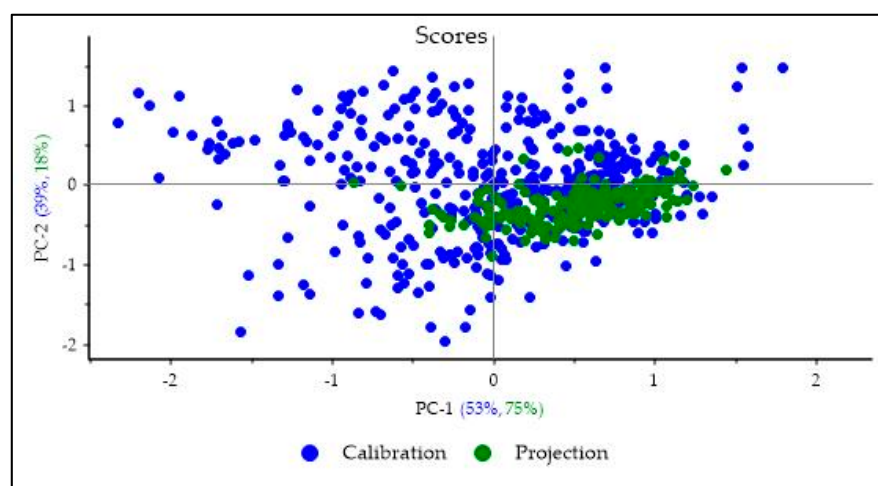


Figure 5. Main component (PC) similarity maps, between the Paraná and target site datasets. Blue scores were obtained by the calibration model using state spectra. Green scores were obtained by the calibration model using local spectra.

Therefore, to obtain success with the use of hybrid curves, the regional, state, or national soil samples that generally make up large spectral libraries must be spectrally very different from the samples collected in new areas where organic matter estimates are of interest, i.e., both datasets must be separated within the occupied spectral space. Thus, after recalibration, this space will be filled by the hybrid curves, forcing the recalibrated model to present better estimation power.

The improvements in organic matter predictions noted after recalibration of the state model were due to the use of the spiking technique without hybridization, as demonstrated by Guerrero et al. [15]. The presence of both datasets within the same spectral domain (Figure 5) led to more positive results, as advocated by Kuang and Mouazen [51] and Nawar and Mouazen [52]. After spiking, a slight improvement in the model fit was noted, especially when assessing the spiked state model using subset four (Figure 6).

However, although improvements in the predictions of the aforementioned attribute were noted, surpassing the results obtained by Nanni et al. [31], who worked with VNIR-SWIR spectroscopy in soil from this region of Paraná, these were lower than expected. With the use of the spiking technique and hybrid spectral curves, the results were expected to be higher than those reported by Daniel et al. [53], Wetterlind et al. [11], and Wang et al. [54], which was not the case.

The recalibration of the state model with selected samples from the target area, despite displaying slightly improved organic matter estimates, indicated that the type of sample directly influences the result. The selected samples should be able to transfer the maximum variability concerning organic matter composition and amounts to the models to be recalibrated, to allow them to adequately estimate this attribute when it is applied to the sample collection area. Within this context, sample selection based on the cluster analysis (strategy four) was the most adequate for state model recalibration.

Another point that should be highlighted refers to the number of soil samples used for recalibration. Only 10 samples were not sufficient to represent the organic matter content distribution of the within the target area, since 59.23% variation was observed. According to Kuang and Mouazen [55], depending on spatial variability, one to two samples per hectare would be sufficient to capture this dissimilarity at a specific site. Nawar and Mouazen [52] suggested the use of four to five samples per ha for recalibration of European spectral libraries in order to provide adequate accuracy in the prediction of soil organic carbon content.

Hong et al. [17] observed improvements in the estimation of soil organic carbon when a larger number of samples (from 10 to 60) were used in the recalibration of the prediction

model. However, it was detected that there were no significant gains when more than 30 samples were used for recalibration. Considering the edaphoclimatic conditions of the studied areas, the researchers suggested selecting from 20 to 30 samples to recalibrate the models to balance the relationship between the modeling cost and predictive accuracy. In turn, Gogé et al. [56] obtained satisfactory results in the estimation of organic carbon applying the spiking technique when they used 50 local samples to recalibrate the prediction model, corresponding to 35% of the total number of samples from the target area.

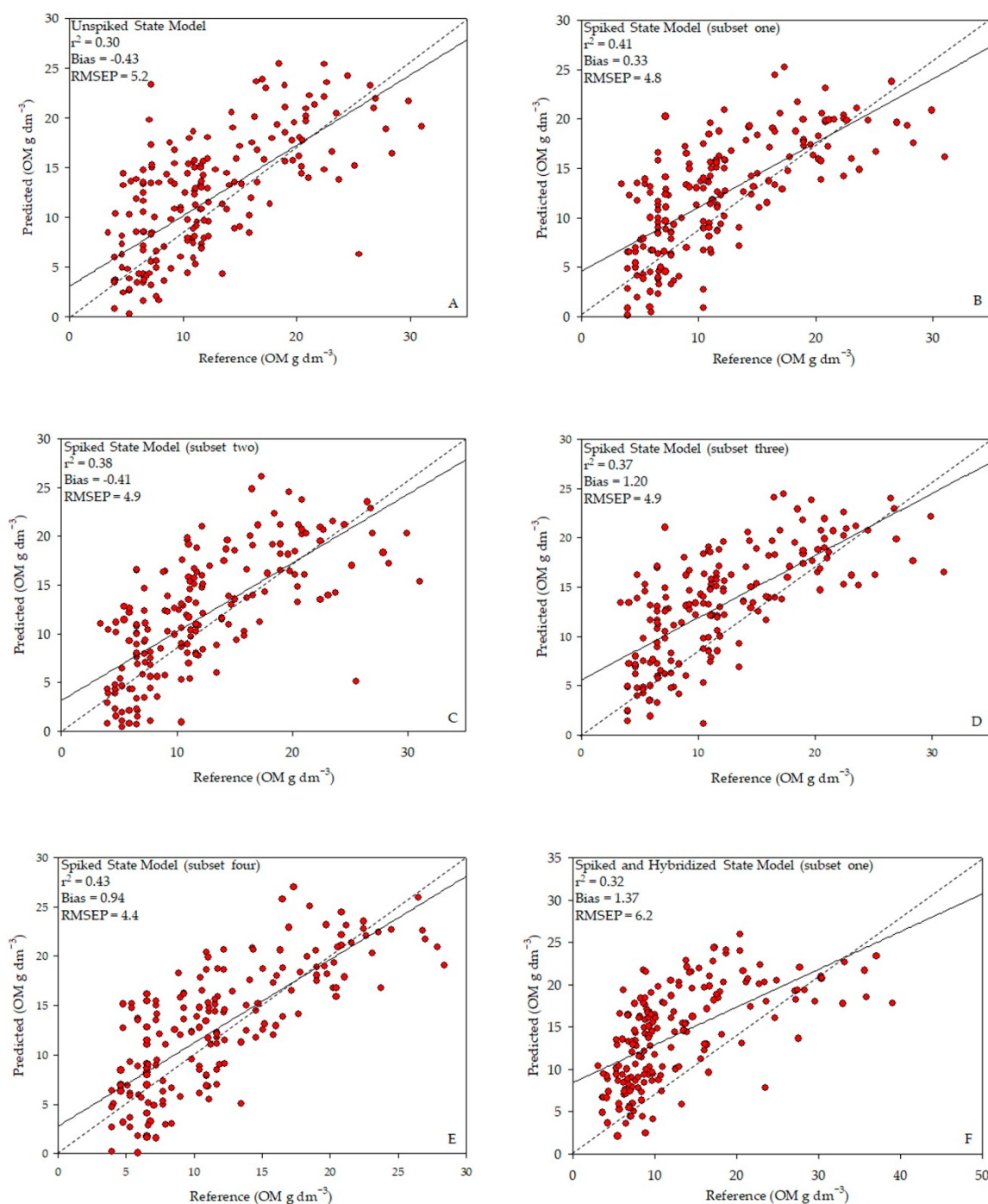


Figure 6. Cont.

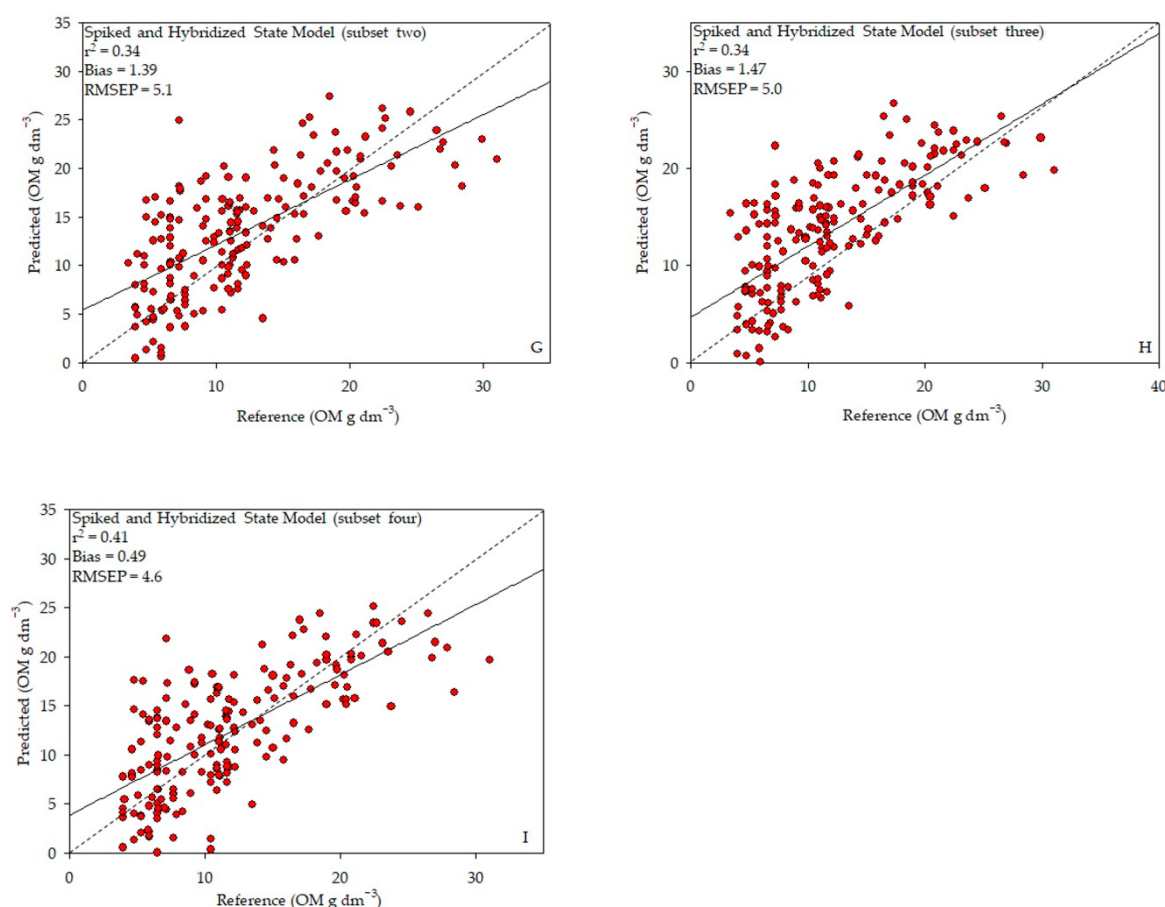


Figure 6. Scatterplots obtained during the prediction phase. (A) Unspiked State Model; (B) Spiked State Model (subset one); (C) Spiked State Model (subset two); (D) Spiked State Model (subset three); (E) Spiked State Model (subset four); (F) Spiked and Hybridized State Model (subset one); (G) Spiked and Hybridized State Model (subset two); (H) Spiked and Hybridized State Model (subset three) and (I) Spiked and Hybridized State Model (subset four). Line 1:1 (dashed); regression line (solid line).

It should be emphasized that increasing the number of samples selected of a target area for recalibration of a prediction model is acceptable to a certain extent, since there is a requirement for analytical results of these samples. If the number is high, spectroscopy ceases to be an attractive technique with innovative potential, becoming an expensive technique when applied to estimation of organic matter or other soil attribute, agreeing with Hong et al. [17].

4. Conclusions

1. The samples selected through a cluster analysis were more effective for state model recalibration, since they were able to transfer more information about the chemical and physical characteristics of the organic matter attribute present in the target area. This fact allowed for a relatively better prediction when compared to the use of other samples selected by other strategies, reaching r^2 and RPD equal to 0.43 and 1.36, respectively, in the spiked state model;

2. The use of hybrid spectral curves did not allow for improvements in organic matter content estimations, since the target area and Paraná state samples occupied the same spectral space. As the hybrid spectrum contains information from both datasets, these will be effective in completing the spectral space if the groups are in different spectral domain, a fact that did not occur in this work;

3. The spiking technique was more effective in state model recalibration than when in conjunction with hybridization, generating more satisfactory results. Maximum RMSEP

and R equal to 4.9 g dm^{-3} against 6.2 g dm^{-3} and 0.76 against 0.71, respectively were observed. The use of selected samples from the target area to recalibrate the state model allowed the incorporation of minimal, but promising, information to improve the estimation of the organic matter content.

4. The study developed in the state of Paraná and in Brazil involving the spiking technique is new, so it is recommended that further research is to be carried out in order to continue the search for the selection of representative samples from new areas, as well as recalibration of predictive models, not only for organic matter, but also for other important soil attributes, which are currently determined by the use of chemical reagents with potential for soil contamination.

Author Contributions: Conceptualization: E.C., M.R.N. and J.A.M.D.; methodology: E.C., L.G.T.C., M.R.N., R.H.F., M.R., R.N.R.S. and J.A.M.D.; formal analysis: L.G.T.C., M.S.C., R.N.R.S., M.R. and G.F.C.S.; data curation: L.G.T.C., M.R.N. and K.M.d.O.; writing—original draft preparation: E.C.; visualization: L.G.T.C. and M.R.N.; project administration: E.C., M.R.N. and L.S.; funding acquisition: E.C., M.R.N. and L.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been funded by the National Council for Scientific and Technological Development, CNPq, through a postdoctoral fellowship grant to the first author and resources to carry out the project analyses foreseen [151658/2012-9]; Coordination of Superior Level Staff Improvement, CAPES; Central Public-Interest Scientific Institution Basal Research Fund [Y2021GH18] and the Talented Young Scientist Program—China Science and Technology Exchange Center [Brazil-19-004].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from the author E.C.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. 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* **2006**, *131*, 59–75. [\[CrossRef\]](#)
2. Sousa Junior, J.G.; Demattê, J.A.M.; Araújo, S.R. Modelos espectrais terrestres e orbitais na determinação de teores de atributos dos solos: Potencial e custos. *Bragantia* **2011**, *70*, 610–621. [\[CrossRef\]](#)
3. Sun, W.; Li, X.; Niu, B. Prediction of soil organic carbon in a coal mining area by VIS-NIR spectroscopy. *PLoS ONE* **2018**, *13*, e0196198. [\[CrossRef\]](#)
4. Lazaar, A.; Mahyou, H.; Gholizadeh, A.; Elhammouti, K.; Bilal, M.; Andich, K.; Roger, J.M.; Monir, A.; Kouotou, D. Potential of VIS-NIR spectroscopy to characterize and discriminate topsoils of different soil types in the Triffa Plain (Morocco). *Soil Sci. Annu.* **2019**, *70*, 54–73. [\[CrossRef\]](#)
5. Janik, L.J.; Merry, R.H.; Skjemstad, J.O. Can mid infra-red diffuse reflectance analysis replace soil extractions? *Aust. J. Exp. Agric.* **1998**, *38*, 681–696. [\[CrossRef\]](#)
6. O'Rourke, S.M.; Holden, N.M. Optical sensing and chemometric analysis of soil organic carbon e a cost effective alternative to conventional laboratory methods? *Soil Use Manag.* **2011**, *27*, 143–155. [\[CrossRef\]](#)
7. Islam, K.; Singh, B.; Mcbratney, A. Simultaneous estimation of several soil properties by ultraviolet, visible, and near-infrared reflectance spectroscopy. *Aust. J. Soil Res.* **2003**, *41*, 1101–1114. [\[CrossRef\]](#)
8. Mcbratney, A.B.; Minasny, B.; Viscarra Rossel, R. Spectral soil analysis and inference systems: A powerful combination for solving the soil data crisis. *Geoderma* **2006**, *136*, 272–278. [\[CrossRef\]](#)
9. Brown, D.J.; Bricklemeyer, R.S.; Millar, P.R. Validation requirements for diffuse reflectance soil characterization models with a case study of VNIR soil C prediction in Montana. *Geoderma* **2005**, *129*, 251–267. [\[CrossRef\]](#)
10. Viscarra Rossel, R.A.; Jeon, Y.S.; Odeh, I.O.A.; McBratney, A.B. Using a legacy soil sample to develop a mid-IR spectral library. *Aust. J. Soil. Res.* **2008**, *46*, 1–16. [\[CrossRef\]](#)
11. Wetterlind, J.; Stenberg, B.; Söderström, M. Increased sample point density in farm soil mapping by local calibration of visible and near infrared prediction models. *Geoderma* **2010**, *156*, 152–160. [\[CrossRef\]](#)
12. Sankey, J.B.; Brown, D.J.; Bernard, M.L.; Lawrence, R.L. Comparing local vs. global visible and near-infrared (VisNIR) diffuse reflectance spectroscopy (DRS) calibrations for the prediction of soil clay, organic C and inorganic C. *Geoderma* **2008**, *148*, 149–158. [\[CrossRef\]](#)

13. D'Acqui, L.P.; Pucci, A.; Janik, L.J. Soil properties of western Mediterranean islands with similar climatic environments by means of mid-infrared diffuse reflectance spectroscopy. *Eur. J. Soil Sci.* **2010**, *61*, 865–876. [\[CrossRef\]](#)
14. Wetterlind, J.; Stenberg, B. Near-infrared spectroscopy for within-field soil characterization: Small local calibrations compared with national libraries spiked with local samples. *Eur. J. Soil Sci.* **2010**, *61*, 823–843. [\[CrossRef\]](#)
15. Guerrero, C.; Stenberg, B.; Wetterlind, J.; Viscarra Rossel, R.A.; Maestre, F.T.; Mouazen, A.M.; Zornoza, R.; Ruiz Sinoga, J.D.; Kuang, B. Assessment of soil organic carbon at local scale with spiked NIR calibrations: Effects of selection and extra-weighting on the spiking subset. *Eur. J. Soil Sci.* **2014**, *65*, 248–263. [\[CrossRef\]](#)
16. Guerrero, C.; Wetterlind, J.; Stenberg, B.; Mouazen, A.M.; Gabarrón-Galeote, M.A.; RuizSinoga, J.D.; Zornoza, R.; Viscarra Rossel, R.A. Do we really need large spectral libraries for local scale SOC assessment with NIR spectroscopy? *Soil Tillage Res.* **2016**, *155*, 501–509. [\[CrossRef\]](#)
17. Hong, Y.; Chen, Y.; Zhang, Y.; Liu, Y.; Liu, Y.; Yu, L.; Liu, Y.; Cheng, H. Transferability of Vis-NIR models for Soil Organic Carbon Estimation between Two Study Areas by using Spiking. *Soil Sci. Soc. Am. J.* **2018**, *82*, 1231–1242. [\[CrossRef\]](#)
18. Guy, A.L.; Siciliano, S.D.; Lamb, E.G. Spiking regional VIS-NIR calibration models with local samples to predict soil organic carbon in two High Arctic polar deserts using a vis-NIR probe. *Can. J. Soil Sci.* **2015**, *95*, 237–249. [\[CrossRef\]](#)
19. Guerrero, C.; Zornoza, R. Extending the geographical validity of NIR models by spiking: Can hybrids spectra act as bridges between sets? A case study for soil organic carbon. Geophysical Research Abstracts; EGU General Assembly: Vienna, Austria, 2011.
20. Hong, Y.; Chen, S.; Zhang, Y.; Chen, Y.; Yu, L.; Liu Yanfang Liu, Y.; Cheng, H.; Liy, Y. Rapid identification of soil organic matter level via visible and near-infrared spec-troscopy: Effects of two-dimensional correlation coefficient and extremelearning machine. *Sci. Total Environ.* **2018**, *644*, 1232–1243. [\[CrossRef\]](#)
21. Memon, M.S.; Guo, J.; Tagar, A.A.; Perveen, N.; Ji, C.; Memon, S.A.; Memon, N. The effects of tillage and straw incorporation on soil organic carbon status, rice crop productivity, and sustainability in the rice-wheat cropping system of eastern China. *Sustainability* **2018**, *10*, 961. [\[CrossRef\]](#)
22. Walkley, A.; Black, I.A. An examination of the Degtjareff method for determining soil organic matter, and proposed modification of the chromic acid titration method. *Soil Sci.* **1934**, *37*, 29–38. [\[CrossRef\]](#)
23. Rosin, N.A.; Dalmolin, R.S.N.; Horst-Heinen, T.Z.; Moura-Bueno, J.M.; da Silva-Sangoi, D.V.; da Silva, L.S. Diffuse reflectance spectroscopy for estimating soil organic carbon and make nitrogen recommendations. *Sci. Agric.* **2020**, *78*, 1–13. [\[CrossRef\]](#)
24. St Luce, M.; Ziadi, N.; Zebarth, B.J.; Grant, C.A.; Tremblay, G.F.; Gregorich, E.G. Rapid determination of soil organic matter quality indicators using visible near infrared reflectance spectroscopy. *Geoderma* **2014**, *232–234*, 449–458. [\[CrossRef\]](#)
25. Empresa Brasileira de Pesquisa Agropecuária. *Mapa de Solos Do Estado Do Paraná*; Embrapa Solos: Rio de Janeiro, Brazil, 2007.
26. Mineropar. *Atlas Geológico Do Estado Do Paraná*; Mineropar: Curitiba, Brazil, 2001; pp. 1–116.
27. World Reference Base for Soil Resources. *International Soil Classification System for Naming Soils and Creating Legends for Soil Maps*, 3rd ed.; FAO: Rome, Italy, 2014; ISBN 978-92-5-108369-7.
28. Empresa Brasileira de Pesquisa Agropecuária. *Manual de Métodos E Análise de Solo*, 2nd ed.; Revista e atualizada; Embrapa: Rio de Janeiro, Brazil, 1997; pp. 1–211.
29. Epiphânio, J.C.N.; Formaggio, A.R.; Valeriano, M.D.M.; Oliveira, J.D. *Comportamento Espectral de Solos Do Estado de São Paulo*, 1st ed.; INPE: São José dos Campos, Brazil, 1992.
30. Labsphere Reflectance Calibration Laboratory. *Spectral Reflectance Target Calibrated from 0.25–2.5 nm Reported in 0.050 nm Intervals*, 1st ed.; Labsphere Inc.: North Sutton, London, UK, 2009.
31. Nanni, M.R.; Cezar, E.; Silva Junior, C.A.D.; Silva, G.F.C.; Gualberto, A.A.S. Partial least squares regression (PLSR) associated with spectral response to predict soil attributes in transitional lithologies. *Arch. Agron. Soil Sci.* **2017**, *64*, 682–695. [\[CrossRef\]](#)
32. Nanni, M.R.; Demattê, J.A.M. Comportamento da linha do solo obtida por espectrorradiometria laboratorial para diferentes classes de solo. *Rev. Bras. Ciênc. Do Solo* **2006**, *30*, 1031–1038. [\[CrossRef\]](#)
33. Lee, K.S.; Sudduth, K.A.; Drummond, T.S.; Lee, D.H.; Kitchen, N.R.; Chung, S.O. Calibration methods for soil property estimation using reflectance spectroscopy. *ASABE* **2010**, *53*, 675–684. [\[CrossRef\]](#)
34. Buddenbaum, H.; Steffens, M. The effects of spectral pretreatments on chemometric analyses of soil profiles using laboratory imaging spectroscopy. *Appl. Environ. Soil Sci.* **2012**, *2012*, 1–12. [\[CrossRef\]](#)
35. Savitzky, A.; Golay, M.J.E. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* **1964**, *36*, 1627–1639. [\[CrossRef\]](#)
36. Cezar, E.; Nanni, M.R.; Guerrero, C.; da Silva Junior, C.A.; Cruciol, L.G.T.; Chicati, M.L.; Silva, G.F.C. Organic matter and sand estimates by spectroradiometry: Strategies for the development of models with applicability at a local scale. *Geoderma* **2019**, *340*, 224–233. [\[CrossRef\]](#)
37. Wrege, M.S.; Steinmetz, S.; Reiser Júnior, C.; de Almeida, I.R. *Atlas Climático da Região Sul do Brasil: Estados do Paraná, Santa Catarina e Rio Grande do Sul*, 2nd ed.; Brasília: Embrapa, Brazil, 2012; pp. 1–331.
38. Martínez, I.; Chervet, A.; Weisskopf, P.; Sturny, W.G.; Etana, A.; Stettler, M.; Forkman, J.; Keller, T. Two decades of no-till in the Oberacker long-term field experiment: Part I. Crop yield, soil organic carbon and nutrient distribution in the soil profile. *Soil Till. Res.* **2016**, *163*, 141–151. [\[CrossRef\]](#)
39. Latz, K.; Weismiller, R.A.; Van Scoyoc, G.E.; Baumgardner, M.F. Characteristic variations in spectral reflectance of selected eroded Alfisols. *Soil Sci. Soc. Am. J.* **1984**, *48*, 1130–1134. [\[CrossRef\]](#)

40. Bowers, S.A.; Hanks, R.J. Reflection of radiant energy from soils. *Soil Sci.* **1965**, *100*, 130–138. [[CrossRef](#)]
41. Vitorrelo, I.; Galvão, L.S. Spectral properties of geologic materials in the 400 to 2500 nm range: Review for applications to mineral exploration and lithologic mapping. *Phot. Int.* **1996**, *2*, 77–96.
42. Demattê, J.A.M.; Garcia, G.J. Alteration of soil properties through a weathering sequence as evaluated by spectral reflectance. *Soil Sci. Soc. Amer. J.* **1999**, *63*, 327. [[CrossRef](#)]
43. Baumgardner, M.F.; Kristof, S.J.; Johannsen, C.J.; Zachary, A.L. Effects of organic matter on the multispectral properties of soils. *Ind. Acad. Sci.* **1969**, *79*, 413–422.
44. Bilgili, A.V.; Van Es, H.M.; Akbas, F.; Durak, A.; Hively, W.D. Visible-near infrared reflectance spectroscopy for assessment of soil properties in a semi-arid area of Turkey. *J. Arid Environ.* **2010**, *74*, 229–238. [[CrossRef](#)]
45. Viscarra Rossel, R.A.; Cattle, S.R.; Ortega, A.; Fouad, Y. In situ measurements of soil colour, mineral composition and clay content by vis–NIR spectroscopy. *Geoderma* **2009**, *150*, 253–266. [[CrossRef](#)]
46. Guerrero, C.; Zornoza, R.; Gómez, I.; Mataix-Beneyto, J. Spiking of NIR regional models using samples from target sites: Effect of model size on prediction accuracy. *Geoderma* **2010**, *158*, 66–77. [[CrossRef](#)]
47. Lazzaretti, B.P.; Da Silva, L.S.; Drescher, G.L.; Dotto, A.C.; Nörnberg, D.B.J.L. Prediction of soil organic matter and clay contents by near-infrared spectroscopy—NIRS. *Cienc. Rural* **2020**, *50*, 1–8. [[CrossRef](#)]
48. Lazaar, A.; Mouazen, A.M.; Hammouti, K.E.; Fullen, M.; Pradhan, B.; Memon, M.S.; Andich, K.; Monir, A. The application of proximal visible and near-infrared spectroscopy to estimate soil organic matter on the triffa plain of Morocco. *Int. Soil Water Conse. Res.* **2020**, *8*, 195–204. [[CrossRef](#)]
49. Qiao, X.X.; Wang, C.; Feng, M.C.; Yang, W.D.; Ding, G.W.; Sun, H.; Liang, Z.Y.; Shi, C.C. Hyperspectral estimation of soil organic matter based on different spectral preprocessing techniques. *Spectrosc. Lett.* **2017**, *50*, 156–163. [[CrossRef](#)]
50. Nawar, S.; Mouazen, A.M. Optimal sample selection for measurement of soil organic carbon using online Vis–NIR spectroscopy. *Comput. Electron. Agric.* **2018**, *151*, 469–477. [[CrossRef](#)]
51. Kuang, B.; Mouazen, A.M. Influence of the number of samples on prediction error of visible and near infrared spectroscopy of selected soil properties at the farm scale. *Eur. J. Soil Sci.* **2012**, *63*, 421–429. [[CrossRef](#)]
52. Nawar, S.; Mouazen, A.M. Predictive performance of mobile vis-near infrared spectroscopy for key soil properties at different geographical scales by using spiking and data mining techniques. *Catena* **2017**, *151*, 118–129. [[CrossRef](#)]
53. Daniel, K.W.; Tripathi, N.K.; Honda, K. Artificial neural network analysis of laboratory and in situ spectra for the estimation of macronutrients in soils of Lop Buri (Thailand). *Aust. J. Soil Res.* **2003**, *41*, 47–59. [[CrossRef](#)]
54. Wang, Y.; Lu, C.; Wang, L.; Song, L.; Wang, R.; Ge, Y. Prediction of Soil Organic Matter Content Using VIS/NIR Soil Sensor. *Sens. Trans. J.* **2014**, *168*, 113–119.
55. Kuang, B.; Mouazen, A.M. Effect of spiking strategy and ratio on calibration of online visible and near infrared soil sensor for measurement in European farms. *Soil Tillage Res.* **2013**, *128*, 125–136. [[CrossRef](#)]
56. Gogé, F.; Gomez, C.; Jolivet, C.; Jofre, R. Which strategy is best to predict soil properties of a local site from a national Vis–NIR database? *Geoderma* **2014**, *213*, 1–9. [[CrossRef](#)]