AI

MDPI

Article

Dimensionality Reduction Statistical Models for Soil Attribute Prediction Based on Raw Spectral Data

Marcelo Chan Fu Wei [1,*], Ricardo Canal Filho [1], Tiago Rodrigues Tavares [1,2], José Paulo Molin [1] and Afrânio Márcio Corrêa Vieira [3,4]

[1] Laboratory of Precision Agriculture (LAP), Department of Biosystems Engineering, "Luiz de Queiroz" College of Agriculture (ESALQ), University of São Paulo (USP), Piracicaba 13418900, Brazil
[2] Center for Nuclear Energy in Agriculture (CENA), University of São Paulo (USP), Piracicaba 13416000, Brazil
[3] Department of Statistics, Federal University of Sao Carlos (UFSCar), São Carlos 13565905, Brazil
[4] Statistics and Agricultural Experimentation Graduate Program, "Luiz de Queiroz" College of Agriculture (ESALQ), University of São Paulo (USP), Piracicaba 13418900, Brazil
* Correspondence: marcelochan@usp.br

Abstract: To obtain a better performance when modeling soil spectral data for attribute prediction, researchers frequently resort to data pretreatment, aiming to reduce noise and highlight the spectral features. Even with the awareness of the existence of dimensionality reduction statistical approaches that can cope with data sparse dimensionality, few studies have explored its applicability in soil sensing. Therefore, this study's objective was to assess the predictive performance of two dimensionality reduction statistical models that are not widespread in the proximal soil sensing community: principal components regression (PCR) and least absolute shrinkage and selection operator (lasso). Here, these two approaches were compared with multiple linear regressions (MLR). All of the modelling strategies were applied without employing pretreatment techniques for soil attribute determination using X-ray fluorescence spectroscopy (XRF) and visible and near-infrared diffuse reflectance spectroscopy (Vis-NIR) data. In addition, the achieved results were compared against the ones reported in the literature that applied pretreatment techniques. The study was carried out with 102 soil samples from two distinct fields. Predictive models were developed for nine chemical and physical soil attributes, using lasso, PCR and MLR. Both Vis-NIR and XRF raw spectral data presented a great performance for soil attribute prediction when modelled with PCR and the lasso method. In general, similar results were found comparing the root mean squared error (RMSE) and coefficient of determination (R^2) from the literature that applied pretreatment techniques and this study. For example, considering base saturation (V%), for Vis-NIR combined with PCR, in this study, RMSE and R^2 values of 10.60 and 0.79 were found compared with 10.38 and 0.80, respectively, in the literature. In addition, looking at potassium (K), XRF associated with lasso yielded an RMSE value of 0.60 and R^2 of 0.92, and in the literature, RMSE and R^2 of 0.53 and 0.95, respectively, were found. The major discrepancy was observed for phosphorus (P) and organic matter (OM) prediction applying PCR in the XRF data, which showed R^2 of 0.33 (for P) and 0.52 (for OM) without using pretreatment techniques in this study, and R^2 of 0.01 (for P) and 0.74 (for OM) when using preprocessing techniques in the literature. These results indicate that data pretreatment can be disposable for predicting some soil attributes when using Vis-NIR and XRF raw data modeled with dimensionality reduction statistical models. Despite this, there is no consensus on the best way to calibrate data, as this seems to be attribute and area specific.

Keywords: multivariate analysis; visible near-infrared; X-ray fluorescence

Citation: Wei, M.C.F.; Canal Filho, R.; Tavares, T.R.; Molin, J.P.; Vieira, A.M.C. Dimensionality Reduction Statistical Models for Soil Attribute Prediction Based on Raw Spectral Data. *AI* **2022**, *3*, 809–819. https://doi.org/10.3390/ai3040049

Academic Editor: Arslan Munir

Received: 6 July 2022
Accepted: 23 September 2022
Published: 30 September 2022

1. Introduction

High spatial density monitoring of soil attributes is a crucial step to build soil maps that can guide better management decisions in crop fields. Fine-scale monitoring of soil

properties is important because if this is neglected, the maps produced are inefficient and unreliable [1], as increasing the spatial density of the soil information directly implies an understanding of the spatial variability of these attributes [2].

The main issues with the traditional soil sampling and analysis methods are cost and time consumption, which become barriers for farmers to increase the density of soil data acquisition [3,4]. Nevertheless, the traditional laboratory analysis consumes reagents that the community aims to diminish, as agriculture production moves towards new global sustainability guidelines. In this scenario, alternative or complementary techniques have slowed down the efforts of soil scientists. Over the last decades, aiming to increase the amount of data related to soil, researchers began to study different soil sensing techniques and their applicability in agriculture [5]. These approaches allow for the rapid acquisition of soil data directly in the field. However, to convert sensor data into agronomic information, predictive models need to be calibrated.

Among the sensing techniques that have shown potential for soil research are the visible and near-infrared diffuse reflectance spectroscopy (Vis-NIR) and X-ray fluorescence spectroscopy (XRF). The core idea for the application of both is to use different spectral data, obtained quickly and without the use of reagents, in order to predict agronomic attributes [6].

The current approach will not replace laboratory analysis, but will allow for augmenting the number of observations, as it uses sensors' output and soil laboratorial analysis to build specific calibrations, transforming spectral data into predictions of physical and chemical soil properties. Therefore, the importance of machine learning (ML) techniques to leverage this novel analysis method is settled. Once ML calibrations are built, soil spectral data can be acquired, allowing for predicting soil attributes in a sustainable, time saving, and cost-effective way [7].

In addition, another highlight of the use of ML calibrations for this task is to create artificial intelligence systems to predict soil attributes. This can reduce human interference in the determination of the physical and chemical soil properties, aiming to simplify and standardize the sample acquisition, processing, and analysis steps [8].

The Vis-NIR spectra of the soil can be related with contents of clay, organic matter (OM), organic carbon (OC), and moisture, and its applications in the literature are observed mainly using laboratory spectral acquisition [9,10]. In the last few years, researchers have developed prediction systems with spectral Vis-NIR acquisition directly in the field, using embedded sensors in agricultural machinery [11–13]. Conversely, XRF is based on the induction of fluorescence in a soil sample through its excitation with an incident X-ray source, as well as the subsequent measurement of specific photons emitted after this process [14]. This technique can accurately measure the total content of some soil elements (Fe, Al, Si, Ca, and K), and can be used to generate indirect calibration models to predict other soil parameters (e.g., cation exchange capacity (CEC), potential of hydrogen (pH), etc.) [15,16].

After data acquisition from the sensors, samples must be taken for chemical and physical analysis in the laboratory to obtain the reference values for each sample in order to fit the predictive models. This last step is usually conducted after applying several combinations of spectra pretreatment methods (e.g., normalization, smoothing, derivative algorithms, and stepwise procedures) aiming to highlight specific features, reduce spectral noise, and select variables to reduce covariates [17,18]. There is not a unique formula to perform this step, nor a method or a set of methods, that have been found to be the best pretreatment sequence. This means that the performance of the prediction will rely on the ability of the researcher who calibrates the models. On the other hand, there are also studies that have applied statistical methods to predict attributes based on spectral data without using pretreatment methods, that have argued that the application of pretreatments can reduce the sensors' predictive performance [19].

In fact, the spectra, whether from Vis-NIR or XRF, have a large number of variables (e.g., $n > 300$), and some of them will deliver a low contribution to prediction models, which, in some cases, can hinder its performance. Nevertheless, it is necessary to consider

the cost of data processing (time consumption), and especially to evaluate its trade-off with performance gain. This is particularly advantageous, because there are methods that can handle the high dimensionality of spectral data without pretreatment or excluding part of the raw data. If a machine learning method can maintain the performance of prediction, using less steps of calibration, or even standardizing the procedure, it can be useful for the soil spectroscopy community. For this purpose, the modeling techniques of principal component regression (PCR) and least absolute shrinkage and selection operator (lasso) regression can be highlighted. These techniques are not widespread in the proximal soil sensing community, and, if successfully applied for predicting the soil attributes, may be an alternative to avoid the spectral preprocessing step. To the best of our knowledge, the performance of these techniques has not yet been evaluated for the prediction of fertility attributes using data from XRF and Vis-NIR sensors, which is the motivation for the execution of this study.

PCR is a regression method that combines principal components analysis with least squares regression [20]. It is a relatively simple, but very useful method [21]. On the other hand, lasso allows for the identification of relevant and irrelevant predictor variables, assigning different weights to each of them [22], a process known as shrinkage or regularization. In the context of proximal soil sensing, although it is easier to find studies applying PCR [18,23] than lasso [24], both methods have not been explored much in the literature and are hardly cited. In this sense, partial least squares (PLS) is often cited as the best method to fit predictive models using soil spectroscopy data, outperforming other models such as artificial neural network (ANN) [25], random forest (RF) [26], or multiple linear regression (MLR) [27]. Therefore, PLS is the most documented technique in the literature [28,29].

Thus, the goal of the present study was to evaluate the predictive performance of the dimensionality reduction statistical models of PCR, lasso, and MLR for soil attribute determination using XRF and Vis-NIR publicly available data without pretreatment, and comparing these results against the ones reported in the literature that applied pretreatment methods.

2. Materials and Methods

2.1. Soil Samples

The dataset used consists of 102 soil samples from the soil database of the Laboratory of Precision Agriculture (LAP) from Luiz de Queiroz College of Agriculture, University of São Paulo [30]. The samples were collected from 0–20 cm depth in two fields under active agricultural production. Field 1 (22°41′57″ S and 47°38′33″ W) is located in the municipality of Piracicaba, State of São Paulo, where 58 samples were collected. Field 2 (14°06′05″ S and 57°45′58″ W) is located in Campo Novo do Parecis, State of Mato Grosso, where the remaining 44 samples were collected. Both fields have considerable textural dissimilarity, as observed from their classification—Lixisol with a clayey texture and Ferralsol with a sandy loam to sandy clay loam texture, respectively. The samples were stored after being air-dried and sieved at 2 mm.

2.2. Physical and Chemical Attributes Analyses

The reference analysis of the clay content, OM, CEC, pH, base saturation (V%), extractable phosphorus (P), extractable potassium (K), extractable calcium (Ca), and extractable magnesium (Mg) were determined in a commercial laboratory for the soil fertility analyses. The laboratorial procedure followed the methodology described by Van Raij et al. [31]. Extractable nutrients (P, K, Ca, and Mg) were determined using ion exchange resin extraction. The OM content was determined via oxidation with a potassium dichromate solution, and the pH was determined via the calcium chloride solution. The texture was determined using the Bouyoucos hydrometer method in a dispersing solution for the clay content.

2.3. Spectral Data Acquisition

The soil spectral data were acquired for all samples using commercial Vis-NIR (351 spectral variables) and XRF (1458 spectral variables) equipment. The Vis-NIR analysis was performed using Veris Vis-NIR spectrometer (Veris Technologies, Salina, KS, USA), which collected the spectra from 343 to 2222 nm, with a spectral resolution of around 5 nm. All of the acquisitions were performed under laboratory conditions, placing the sample at a circular sapphire window located in the bottom portion of a shank module. This spectrometer self-calibrated before each spectra acquisition by collecting a dark reference measurement and a known internal reference material measurement. Spectral regions at 343–432 and 2153–2222 nm were removed due to the high presence of noise, resulting in a raw spectra from 437 to 2149 nm. For the XRF spectra acquisition, the portable XRF equipment Tracer III-SD (Bruker AXS, Madison, WI, USA) was used. The equipment was configured using the instrumental conditions suggested by Tavares et al. [32]. Samples were scanned in triplicate and then averaged for further analysis.

2.4. Data Modeling

For all of the three methods, data from the two sensors were individually used. All of the processes were conducted in RStudio [33]. Raw spectral data were used as the input variables to fit the prediction models of the clay, OM, CEC, pH, V%, P, K, Ca, and Mg. The prediction models were fitted by applying three regression models: MLR, PCR, and lasso regression.

The MLR method is a regression analysis that has one target related to more than one feature, where the target can be estimated by Equation (1) [34].

$$\mathrm{Y} = \mathrm{X}\beta + e \tag{1}$$

where Y is a ($n \times 1$) target vector, X is a ($n \times p$) features matrix (predictor variables), β is a $p \times 1$ vector of unknown coefficients, and e is a $n \times 1$ random vector of errors.

The PCR method occurs in three steps: (a) perform principal component analysis on the observed data matrix to obtain principal components; (b) apply linear regression to obtain the vector of estimated regression coefficients; and (c) use PCA loadings (eigenvectors) to obtain the PCR estimator (β), as shown in Equation (2).

$$\widehat{\beta_k} = V_k \widehat{\delta_k} \tag{2}$$

where $\widehat{\beta}$ is the PCR estimator, k belongs to {1, . . . , p}, p is the number of covariates, V is the orthonormal set of eigenvectors, and $\widehat{\delta}$ is the vector of estimated regression coefficients.

Lasso is a regression with an l1-norm penalty aiming to find β = {βj}, which minimizes Equation (3) [35].

$$\sum_{i=1}^{N} (y_i - \sum_j x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \left|\beta_j\right| \tag{3}$$

where x_{ij} is the standardized features, y_i is the centered target values, and i = 1, 2, . . . , N and j = 1, 2, . . . , p.

Data were split into 75% as the training set and 25% as the validation set. Therefore, 76 samples were used for training and 26 for validation. The data split was randomly selected using the seed (666). Although this is a random procedure to avoid bias, the seed allows for repeatability after the process is finished. Both PCR and lasso were performed using the R language library caret [36], setting the method as pcr() for PCR and glmnet() for lasso.

In PCR, the optimal number of principal components was defined using the tuning control function trainControl with the cross-validation (cv) as 10 and the tuneLength (number of principal components tested) equal to 30 in the training dataset. The elbow rule [37] was used to minimize the root mean squared error (RMSE) of cross-validation and to maximize the variance explained, aiming, when possible, to obtain at least 70% of

the explained variance. Although this rule will not always choose the model that present the highest coefficient of determination (R^2) and lowest RMSE, it follows the principle of parsimony in multivariate calibration, assuming that of the two models with meaningful predictions, the one with fewer parameters is preferred [38]. The scale within train function was set to TRUE as it standardized each variable before the generation of the principal components. Then, the model was tuned and applied to the validation dataset, calculating the RMSE and R^2 of the prediction.

For the lasso method, standardize within the function train was set to TRUE for data scaling, which removes the effect of features that present different unit/magnitude. The best alpha (α) and lambda (λ) values were extracted from the fitted model, which was applied to the validation dataset and obtained the RMSE and R^2 of prediction.

MLR method was applied using the function lm. The model was fitted on the training dataset and then applied to the validation dataset, obtaining the respective RMSE and R^2 for each attribute predicted.

3. Results and Discussion

For the PCR prediction, both Vis-NIR and XRF soil attributes prediction presented R^2 values from 0.52 to 0.85, being P the only exception (Table 1). The R^2 for P was 0.03 and 0.33 for Vis-NIR and XRF, respectively, indicating that P was better predicted, for the data used in this study, using PCR prediction and XRF sensor technique. P prediction yielded poor parameters in both studies because it has no direct spectral response in Vis-NIR [9] nor emission lines for XRF technique [32]. The accurate prediction of P, sometimes observed in literature [39,40], lies into the covariation with other soil properties of the studied area, allowing indirect calibrations [41].

Table 1. Results from the principal component regression (PCR) fitting model applied to the soil physical-chemical attributes studied.

Sensor	Target Variable	Number of Components Used	Variance Explained (%)	Validation		Tavares et al. [42]	
				RMSE	R^2	RMSE	R^2
Vis-NIR	Clay	2	84.91	42.80	0.80	27.32	0.93
	OM	7	72.29	2.95	0.72	2.10	0.86
	CEC	10	71.39	18.91	0.56	18.66	0.51
	pH	19	70.47	0.28	0.58	0.34	0.19
	V%	2	75.99	10.60	0.79	10.38	0.80
	P	3	11.87	13.76	0.03	12.05	0.07
	K	4	66.29	0.90	0.82	1.20	0.74
	Ca	2	69.61	12.80	0.63	10.98	0.68
	Mg	10	71.05	9.34	0.59	8.85	0.52
XRF	Clay	1	83.00	45.63	0.78	29.40	0.92
	OM	19	66.28	3.89	0.52	3.01	0.74
	CEC	4	72.46	15.37	0.70	10.19	0.88
	pH	13	53.80	0.24	0.68	0.33	0.34
	V%	1	78.87	9.68	0.83	5.60	0.95
	P	15	41.26	11.39	0.33	13.27	0.01
	K	2	67.47	0.82	0.85	0.53	0.95
	Ca	1	72.06	11.99	0.68	4.09	0.96
	Mg	4	71.54	8.13	0.68	4.28	0.89

Vis-NIR: visible and near-infrared diffuse reflectance spectroscopy; XRF: X-ray fluorescence spectroscopy; Clay (g kg^{-1}); OM: organic matter content (g kg^{-1}); CEC: cation exchange capacity ($mmol_c$ kg^{-1}); pH: potential of hydrogen; V%: base saturation ($mmol_c$ kg^{-1}); P: phosphorus ($mmol_c$ kg^{-1}); K: potassium ($mmol_c$ kg^{-1}); Ca: calcium ($mmol_c$ kg^{-1}); Mg: magnesium ($mmol_c$ kg^{-1}); RMSE: root mean squared error; R^2: coefficient of determination.

The application of PCR on XRF raw data used less components when compared with the Vis-NIR raw data, except for OM, pH, and P. It is well known that clay and organic matter have a direct spectral response on Vis-NIR [41]. Therefore, the model can maximize the explained variance with a low number of principal components (PC), as it will identify well established spectral regions that are modified due to the present amount

of these attributes [9]. In this context, other soil attributes that present a strong linear correlation with clay and OM (i.e., Vis-NIR primary response attributes) can be indirectly predicted [41]. Once this covariation exists, the model will tend to identify the same spectral regions used for the primary response attributes. However, the portion of the variance that are not correlated with primary response attributes will be randomly assigned by the model. Hence, the weaker the linear correlation, the greater the number of PCs the model will need to increase the explained variance.

The indirect calibrations built with PCR (i.e., CEC, pH, V%, K, Ca, and Mg) presented a similar or slightly improved performance in comparison with the results in Tavares et al. [42], although this applied the elbow rule. In this sense, the pH prediction of this study was highlighted due to the R^2 of 0.58 (Vis-NIR) and 0.68 (XRF) obtained by PCR. These results reinforced the capacity of dimensionality reduction models for identifying important pH-related spectral regions as those reported in the literature by wavebands of around 2200 nm of O–H and N–H bonds, for OM direct responses that can also be indirectly related to pH, as described by Chang et al. [43] and Li et al. [44].

The different calibration strategies used in Tavares et al. [42] achieved slightly better performances than the results reached by the present study. For example, the R^2 for clay and OM obtained by the authors were 0.93 and 0.86, respectively, compared with 0.80 and 0.72 achieved by our study. Despite this difference, the method used in this study was justified by the principle of parsimony applied to the multivariate calibration [38], reducing the calibration parameters and providing satisfactory predictions, even diminishing the performance.

Comparing the lasso predictions obtained in this study with the results reported in Tavares et al. [42] (Table 2), a similar pattern as that reported for PCR predictions was noted. Despite the sensor used, clay and OM had their prediction performance reduced, but still presented a satisfactory prediction. The other calibrations presented a similar performance, except for the CEC, Ca, and Mg prediction from the XRF data, for which the RMSE from lasso increased by over 50% when compared with the results from the above-mentioned authors.

Table 2. Results from least absolute shrinkage and selection operator (lasso) regression fitting models applied to the soil physical-chemical attributes studied.

Sensor	Target Variable	α	λ	Validation		Tavares et al. [42]	
				RMSE	R^2	RMSE	R^2
Vis-NIR	Clay	0.10	5.140	38.03	0.84	27.32	0.93
	OM	0.10	0.218	2.95	0.71	2.10	0.86
	CEC	0.10	1.149	21.27	0.44	18.66	0.51
	pH	0.55	0.044	0.35	0.33	0.34	0.19
	V%	0.10	1.140	10.98	0.78	10.38	0.80
	P	1.00	3.220	13.30	0.06	12.05	0.07
	K	0.10	0.387	0.94	0.81	1.20	0.74
	Ca	0.55	0.929	12.95	0.62	10.98	0.68
	Mg	0.10	0.557	10.72	0.45	8.85	0.52
XRF	Clay	0.10	5.554	42.23	0.81	29.40	0.92
	OM	0.10	2.513	3.58	0.59	3.01	0.74
	CEC	0.55	4.329	16.45	0.68	10.19	0.88
	pH	0.10	0.148	0.24	0.70	0.33	0.34
	V%	0.10	4.051	4.63	0.96	5.60	0.95
	P	1.00	4.432	12.42	0.34	13.27	0.01
	K	0.10	0.145	0.60	0.92	0.53	0.95
	Ca	1.00	1.110	7.91	0.86	4.09	0.96
	Mg	0.10	0.660	6.63	0.81	4.28	0.89

Vis-NIR: visible and near-infrared diffuse reflectance spectroscopy; XRF: X-ray fluorescence spectroscopy; Clay (g kg^{-1}); OM: organic matter content (g kg^{-1}); CEC: cation exchange capacity ($mmol_c$ kg^{-1}); pH: potential of hydrogen; V%: base saturation ($mmol_c$ kg^{-1}); P: phosphorus ($mmol_c$ kg^{-1}); K: potassium ($mmol_c$ kg^{-1}); Ca: calcium ($mmol_c$ kg^{-1}); Mg: magnesium ($mmol_c$ kg^{-1}); RMSE: root mean squared error; R^2: coefficient of determination.

Considering Vis-NIR, the P predictions were not accurately predicted in this study (RMSE and R^2 of 13.30 $mmol_c$ kg^{-1} and 0.06, respectively) nor in that of Tavares et al. [42] (RMSE and R^2 of 12.05 $mmol_c$ kg^{-1} and 0.07, respectively). Both Vis-NIR and XRF predictions of V% presented low RMSE (Vis-NIR using lasso achieved 10.98 $mmol_c$ kg^{-1} and Tavares et al. [42] reached 10.38 $mmol_c$ kg^{-1}; XRF using lasso obtained 4.63 $mmol_c$ kg^{-1} and Tavares et al. [42] achieved 5.60 $mmol_c$ kg^{-1}) and high R^2 (Vis-NIR using lasso achieved 0.78 and Tavares et al. [42] reached 0.80; XRF using lasso obtained 0.96 and Tavares et al. [42] achieved 0.95) values. This comparison highlights that instead of using pretreatment techniques, which can be time consuming, dimensionality reduction statistical models are capable of coping with soil spectral data for the successful predictions of the fertility attributes.

In a study of moisture sensibility of soil attributes prediction using Vis-NIR spectra [45], the authors applied the average pretreatment (averaging values of the spectra within a given interval to reduce dimensionality) before fitting partial least squares regression (PLSR), random forest (RF) regression, artificial neural network (ANN), and support vector machine (SVM) models. As a result, no major discrepancy in prediction among the different models tested was found. This corroborates the difficulty in defining whether pretreatment methods can regularly cope with a singular statistical model, but not with others, as no pattern was observed in the results. This emphasizes the complexity to settle a standard procedure to build ML calibrations for the soil spectral data using pretreatment techniques.

Another study used Vis-NIR spectra to predict the soil attributes, which aimed to define the lime requirement doses [46]. The pretreatments involved smoothing and resampling for all of the data, and then the authors tested the maximum normalization, multiplicative scatter correction, and standard normal variate to build PLSR predictive models. The best results found for RMSE were 0.33 for pH. Compared with the results reported in this study, applying lasso (RMSE values of 0.35 for pH) and PCR (RMSE values of 0.28 for pH), no major differences were observed, even compared with the results in Tavares et al. [42], which presented RMSE values of 0.34.

More recently, in a study using Vis-NIR spectra, testing several pretreatment sequences, the authors used the strategy of choosing a specific combination for each individual soil property, aiming to predict using PLSR [39]. Comparing the results of the prediction from the models that contained only laboratory measured spectra, the pH presented R^2 of 0.45 and RMSE of 0.56. The R^2 values for K, Ca, and Mg were 0.13, 0.48, and 0.25, respectively. Thus, we can infer that the results in this study, without pretreatment and using alternative statistical modelling presented a higher R^2, indicating that dimensionality reduction statistical models can be used to predict soil attributes, with the advantage of not requiring spectral preprocessing.

Another study evaluated alternative statistical approaches (RF regression, and SVM with radial and linear kernel) to predict the soil physical attributes using both Vis-NIR and XRF spectra, separately and in tandem [47]. The authors tested the datasets with and without pretreatment. Analyzing the results, just as observed in the present study, there was also no major difference between the predictions, and no pattern was found. The main factors of variation appeared to be the technique used (Vis-NIR, XRF, or the combination of both) and the soil sampling depth.

These comparisons corroborate and incite the question of whether the use of pretreatment methods is decisive or not for soil attribute prediction using spectral data, as there are few methods that automatically apply pretreatment techniques and fit predictive models, such as the 'all-possibilities' approach (APA) of Kopačková et al. [48]. The APA method is an algorithm that automatically fits predictive models (e.g., ANN and PLSR) using several pretreatment methods, including averaging, centering, smoothing, standardization, normalization, and transformations, among others. Furthermore, for the application of this algorithm, it is necessary to have a high computational processing power. In their study, the PARACUDA® computing engine was used, which the authors defined as "an extremely computer power-consuming method and thus it runs on a grid based supercomputer with many processing cores for rapid analysis". Without automatic methods, the majority of

studies that apply pretreatment techniques on soil spectral data rely on the expertise of the user to calibrate the predictive models. Therefore, the use of alternative approaches is an opportunity to standardize the procedure, and to reduce the time and cost consumption.

Depending on the purpose of the study, pretreatment is indeed needed for some applications, e.g., detailed chemometrics analysis, where the visualization of the spectra range is important for a given attribute. On the other hand, looking at only predictive models based on spectral data, there may be an opportunity to explore the application of predictive models based on raw soil spectral data, as observed in a study that applied deep learning on raw spectral data to predict fresh fruit attributes [19].

Further in this subject, Velliangiri and Alagumuthukrishnan [49] described that using dimensionality reduction models, such as PCR and lasso, can aid in the removal of noisy, redundant, and irrelevant data, which are similar characteristics to the goals of pretreatment methods. This corroborates the methods tested in this study and also explains the results that were found by the models calibrated using PCR and lasso

The results for the MLR fitted models (Table 3) indicate that the use of raw spectral data to calibrate MLR models for soil attribute prediction is not feasible independently of the source of the data tested (Vis-NIR or XRF). This is because MLR is a method that handles high-dimensional data poorly in comparison with PCR and lasso. The assumptions underlying the dimensionality reduction statistical models include the correlations among the predictors, the noise to signal, and model errors [49]. This is often more realistic than the MLR assumptions of independent and error free predictors, allowing these models to handle high-dimensional data better than ordinary MLR [27].

Table 3. The results from the multiple linear regression fitting models applied to the soil physical-chemical attributes studied.

Sensor	Target Variable	Validation		Tavares et al. [42]	
		RMSE	R^2	RMSE	R^2
Vis-NIR	Clay	158.60	0.32	27.32	0.93
	OM	38.98	0.01	2.10	0.86
	CEC	86.61	0.10	18.66	0.51
	pH	1.37	0.01	0.34	0.19
	V%	81.66	0.15	10.38	0.80
	P	70.10	0.03	12.05	0.07
	K	9.54	0.22	1.20	0.74
	Ca	83.37	0.07	10.98	0.68
	Mg	35.72	0.09	8.85	0.52
XRF	Clay	448.21	0.02	29.40	0.92
	OM	33.79	0.13	3.01	0.74
	CEC	444.65	0.01	10.19	0.88
	pH	10.20	0.01	0.33	0.34
	V%	482.51	0.00	5.60	0.95
	P	51.07	0.09	13.27	0.01
	K	23.35	0.01	0.53	0.95
	Ca	394.45	0.00	4.09	0.96
	Mg	314.98	0.03	4.28	0.89

Vis-NIR: visible and near-infrared diffuse reflectance spectroscopy; XRF: X-ray fluorescence spectroscopy; Clay (g kg^{-1}); OM: organic matter content (g kg^{-1}); CEC: cation exchange capacity ($mmol_c$ kg^{-1}); pH: potential of hydrogen; V%: base saturation ($mmol_c$ kg^{-1}); P: Phosphorus ($mmol_c$ kg^{-1}); K: potassium ($mmol_c$ kg^{-1}); Ca: calcium ($mmol_c$ kg^{-1}); Mg: magnesium ($mmol_c$ kg^{-1}); RMSE: root mean squared error; R^2: coefficient of determination.

Note that pretreatment followed by MLR improved the prediction accuracy, as shown in the results from Tavares et al. [42], highlighting that MLR can be applied to predict soil attributes based on spectral data. However, the application of pretreatment techniques is necessary.

Despite the close values of the quality indicators of the dimensionality reduction statistical models and the models using pretreatment methods, it was not possible to observe an optimal method that is unanimously the best for all attributes and sensors. It is important to highlight that the aim of this study was not to indicate the best set (sensor and method) to predict soil physical-chemical attributes, but to show the potential of applying unusual methods that can be suitable to fit predictive models based on the soil raw spectral data.

Spectra pretreatment relies on the ability of the person who calibrates the model, to choose the combination of techniques that will extract the best prediction. The proposed method in this study is advantageous as it presents the potential for standardizing the model calibration step, thus diminishing human interference.

The sensors used in the laboratory to obtain spectral data are being adapted to be used directly in the field, with Vis-NIR being the only that has already been tested. The main use in the laboratory justifies the appliance of pretreatment techniques before fitting the predictive model. However, looking at the advances in agricultural machines towards proximal soil sensing, strategies to predict soil attributes online in a standard manner, and reducing the time and processing cost needed as much as possible, can be a helpful way to leverage soil sensing techniques with a high spatial resolution [3]. Therefore, the advances in this research area can further decide whether methods using less processing steps become more interesting than the classical approaches. The comparison of the results found in this study was restricted to laboratory measured spectra, as there was a disadvantage in the quality of online acquired spectra when compared with the laboratory acquisition [41].

Future works should evaluate different statistical approaches, including those conducted in this study, comparing the results from models built with and without pretreatment techniques for online and laboratory spectra acquisition. If possible, they should be applied in a dataset that presents a large variability and number of samples. In addition, they should be postulated not only by metrics that assess mean values, such as RMSE, but also by the correlation among predicted values, so as to observe the discrepancy among the different modelling methods.

4. Conclusions

The application of principal component regression and the lasso method overperformed the multiple linear regression approach. The R^2 values found in this study applying PCR and the lasso method ranged from 0.33 to 0.96 and 0.03 to 0.84 for XRF and Vis-NIR sensor data, respectively. The MLR approach did not present satisfactory results for any of the evaluated attributes ($R^2 \leq 0.32$), indicating that it cannot be applied on raw spectral data. When comparing the results achieved in the present study using the tested dimensionality reduction statistical models with the studies in the literature that applied data pretreatment methods, we noticed that both strategies presented similarly satisfactory results. Hence, this suggests the possibility of predicting soil attributes without applying pretreatment methods, making the data processing faster.

Author Contributions: Conceptualization, M.C.F.W. and R.C.F.; methodology, M.C.F.W., R.C.F. and A.M.C.V.; formal analysis, M.C.F.W., R.C.F. and A.M.C.V.; writing—review and editing, M.C.F.W., R.C.F., T.R.T., J.P.M. and A.M.C.V. All authors have read and agreed to the published version of the manuscript.

Funding: Soil fertility tests were funded by CNPq, "Edital de Chamada Universal", grant number 458180/2014-9. Tiago R. Tavares was funded by the São Paulo Research Foundation (FAPESP), Grant No. 2020/16670-9.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is available at: Ref. [30].

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

References

1. Wollenhaupt, N.C.; Wolkowski, R.P.; Clayton, M.K. Mapping Soil Test Phosphorus and Potassium for Variable-Rate Fertilizer Application. *J. Prod. Agric.* **1994**, *7*, 441–448. [CrossRef]
2. Cherubin, M.R.; Santi, A.L.; Eitelwein, M.T.; Amado, T.J.C.; Simon, D.H.; Damian, J.M. Dimensão da malha amostral para caracterização da variabilidade espacial de fósforo e potássio em Latossolo Vermelho. *Pesqui. Agropecuária Bras.* **2015**, *50*, 168–177. (In Portuguese) [CrossRef]
3. Viscarra Rossel, R.A.; Adamchuk, V.I.; Sudduth, K.A.; McKenzie, N.J.; Lobsey, C. Proximal soil sensing: An effective approach for soil measurements in space and time. In *Advances in Agronomy*; Sparks, D.L., Ed.; Academic Press: London, UK, 2011; pp. 243–291.
4. Wang, D.; Chakraborty, S.; Weindorf, D.C.; Li, B.; Sharma, A.; Paul, S.; Ali, N. Synthesized use of VisNIR DRS and PXRF for soil characterization: Total carbon and total nitrogen. *Geoderma* **2015**, *243–244*, 157–167. [CrossRef]
5. Molin, J.P.; Tavares, T.R. Sensor Systems for Mapping Soil Fertility Attributes: Challenges, Advances, and Perspectives in Brazilian Tropical Soils. *Eng. Agrícola* **2019**, *39*, 126–147. [CrossRef]
6. Kuang, B.; Mahmood, H.S.; Quraishi, M.Z.; Hoogmoed, W.B.; Mouazen, A.M.; van Henten, E.J. Sensing soil properties in the laboratory, in situ, and on-line: A review. In *Advances in Agronomy*; Sparks, D.L., Ed.; Academic Press: London, UK, 2012; pp. 155–223.
7. Sharma, R.; Kamble, S.S.; Gunasekaran, A.; Kumar, V.; Kumar, A. A systematic literature review on machine learning applications for sustainable agriculture supply chain performance. *Comput. Oper. Res.* **2020**, *119*, 104926. [CrossRef]
8. Rossi, P.; Mangiavacchi, P.L.; Monarca, D.; Cecchini, M. Smart Machinery and Devices for Reducing Risks from Human-Machine Interference in Agriculture: A Review. In *Safety, Health and Welfare in Agriculture and Agro-food Systems*; Biocca, M., Cavallo, E., Cecchine, M., Failla, S., Romano, E., Eds.; Springer: Cham, Switzerland, 2022; pp. 195–204.
9. Nocita, M.; Stevens, A.; van Wesemael, B.; Aitkenhead, M.; Bachmann, M.; Barthès, B.; Dor, E.B.; Brown, D.J.; Clairotte, M.; Csorba, A.; et al. Soil spectroscopy: An alternative to wet chemistry for soil monitoring. In *Advances in Agronomy*; Sparks, D.L., Ed.; Academic Press: London, UK, 2015; pp. 139–159.
10. Pasquini, C. Near infrared spectroscopy: A mature analytical technique with new perspectives–A review. *Anal. Chim. Acta* **2018**, *1026*, 8–36. [CrossRef]
11. Bönecke, E.; Meyer, S.; Vogel, S.; Schröter, I.; Gebbers, R.; Kling, C.; Kramer, E.; Lück, K.; Nagel, A.; Philipp, G.; et al. Guidelines for precise lime management based on high-resolution soil pH, texture and SOM maps generated from proximal soil sensing data. *Precis. Agric.* **2020**, *22*, 493–523. [CrossRef]
12. Vogel, S.; Bönecke, E.; Kling, C.; Kramer, E.; Lück, K.; Philipp, G.; Rühlmann, J.; Schröter, I.; Gebbers, R. Direct prediction of site-specific lime requirement of arable fields using the base neutralizing capacity and a multi-sensor platform for on-the-go soil mapping. *Precis. Agric.* **2022**, *23*, 127–149. [CrossRef]
13. Munnaf, M.A.; Haesaert, G.; Van Meirvenne, M.; Mouazen, A.M. Multi-sensors data fusion approach for site-specific seeding of consumption and seed potato production. *Precis. Agric.* **2021**, *22*, 1890–1917. [CrossRef]
14. Kalnicky, D.J.; Singhvi, R. Field portable XRF analysis of environmental samples. *J. Hazard. Mater.* **2001**, *83*, 93–122. [CrossRef]
15. Nawar, S.; Delbecque, N.; Declercq, Y.; De Smedt, P.; Finke, P.; Verdoodt, A.; Mouazen, A.M. Can spectral analyses improve measurement of key soil fertility parameters with X-ray fluorescence spectrometry? *Geoderma* **2019**, *350*, 29–39. [CrossRef]
16. Tavares, T.R.; Molin, J.P.; Javadi, S.H.; De Carvalho, H.W.P.; Mouazen, A.M. Combined Use of Vis-NIR and XRF Sensors for Tropical Soil Fertility Analysis: Assessing Different Data Fusion Approaches. *Sensors* **2021**, *21*, 148. [CrossRef] [PubMed]
17. Munnaf, M.A.; Guerrero, A.; Nawar, S.; Haesaert, G.; Van Meirvenne, M.; Mouazen, A.M. A combined data mining approach for on-line prediction of key soil quality indicators by Vis-NIR spectroscopy. *Soil Tillage Res.* **2021**, *205*, 104808. [CrossRef]
18. Javadi, S.H.; Munnaf, M.A.; Mouazen, A.M. Fusion of Vis-NIR and XRF spectra for estimation of key soil attributes. *Geoderma* **2021**, *385*, 114851. [CrossRef]
19. Mishra, P.; Rutledge, D.N.; Roger, J.-M.; Wali, K.; Khan, H.A. Chemometric pre-processing can negatively affect the performance of near-infrared spectroscopy models for fruit quality prediction. *Talanta* **2021**, *229*, 122303. [CrossRef] [PubMed]
20. Jolliffe, I. *Principal Component Analysis (pp. 1094–1096). Springer Berlin Heidelberg. RESUME SELİN DEĞİRMECİ Marmara University, Goztepe Campus ProQuest Number: ProQuest). Copyright of the Dissertation Is Held by the Author. All Rights Reserved, 28243034*; Springer: New York, NY, USA, 2011.
21. Agarwal, A.; Shah, D.; Shen, D.; Song, D. On robustness of principal component regression. *J. Am. Stat. Assoc.* **2021**, *10*, 1–34. [CrossRef]
22. Lee, J.H.; Shi, Z.; Gao, Z. On LASSO for predictive regression. *J. Econ.* **2021**, *229*, 322–349. [CrossRef]
23. Pudełko, A.; Chodak, M. Estimation of total nitrogen and organic carbon contents in mine soils with NIR reflectance spectroscopy and various chemometric methods. *Geoderma* **2020**, *368*, 114306. [CrossRef]
24. Bricklemyer, R.S.; Brown, D.J.; Turk, P.J.; Clegg, S. Comparing vis-NIRS, LIBS, and combined vis-NIRS-LIBS for intact soil core soil carbon measurement. *Soil Sci. Soc. Am. J.* **2018**, *82*, 1482–1496. [CrossRef]
25. Kuang, B.; Tekin, Y.; Mouazen, A.M. Comparison between artificial neural network and partial least squares for on-line visible and near infrared spectroscopy measurement of soil organic carbon, pH and clay content. *Soil Tillage Res.* **2015**, *146*, 243–252. [CrossRef]

26. Knox, N.; Grunwald, S.; McDowell, M.; Bruland, G.; Myers, D.; Harris, W. Modelling soil carbon fractions with visible near-infrared (VNIR) and mid-infrared (MIR) spectroscopy. *Geoderma* **2015**, *239–240*, 229–239. [CrossRef]
27. Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130. [CrossRef]
28. 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.* **2011**, *43*, 1398–1410. [CrossRef]
29. Rossel, R.V.; Behrens, T. Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma* **2010**, *158*, 46–54. [CrossRef]
30. Tavares, T.R.; Molin, J.P.; Nunes, L.C.; Alves, E.E.N.; Krug, F.J.; de Carvalho, H.W.P. Spectral data of tropical soils using dry-chemistry techniques (Vis-NIR, XRF, and LIBS): A dataset for soil fertility prediction. *Data Brief* **2022**, *41*, 108004. [CrossRef]
31. Van Raij, B.; Andrade, J.C.; Cantarela, H.; Quaggio, J.A. *Análise Química Para Avaliação de Solos Tropicais*; IAC: Campinas, Brazil, 2001; 285p. (In Portuguese)
32. Tavares, T.R.; Molin, J.P.; Nunes, L.C.; Alves, E.E.N.; Melquiades, F.L.; de Carvalho, H.W.P.; Mouazen, A.M. Effect of X-Ray Tube Configuration on Measurement of Key Soil Fertility Attributes with XRF. *Remote Sens.* **2020**, *12*, 963. [CrossRef]
33. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2018.
34. Olive, D.J. *Linear Regression*, 1st ed.; Springer: Cham, Switzerland, 2017; Multiple linear regression; pp. 17–83.
35. Tibshirani, R. Regression Shrinkage and Selection via the Lasso: A retrospective. *J. R. Stat. Soc.* **2011**, *73*, 267–288. [CrossRef]
36. Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Softw.* **2008**, *28*, 1548–7660. [CrossRef]
37. Tracy, T.; Fu, Y.; Roy, I.; Jonas, E.; Glendenning, P. *High Performance Computing*; Kunkel, J., Balaji, P., Dongarra, J., Eds.; Springer: Cham, Switzerland, 2015; pp. 200–218.
38. Seasholtz, M.B.; Kowalski, B. The parsimony principle applied to multivariate calibration. *Anal. Chim. Acta* **1993**, *277*, 165–177. [CrossRef]
39. Abdul Munnaf, M.; Nawar, S.; Mouazen, A.M. Estimation of Secondary Soil Properties by Fusion of Laboratory and On-Line Measured Vis–NIR Spectra. *Remote. Sens.* **2019**, *11*, 2819. [CrossRef]
40. Mouazen, A.M.; Kuang, B. On-line visible and near infrared spectroscopy for in-field phosphorous management. *Soil Tillage Res.* **2016**, *155*, 471–477. [CrossRef]
41. Stenberg, B.; Viscarra-Rossel, R.A.; Mouazen, A.M.; Wetterlind, J. Visible and near infrared spectroscopy in soil science. In *Advances in Agronomy*; Sparks, D.L., Ed.; Academic Press: London, UK, 2010; pp. 163–215.
42. Tavares, T.; Molin, J.; Nunes, L.; Wei, M.; Krug, F.; de Carvalho, H.; Mouazen, A. Multi-Sensor Approach for Tropical Soil Fertility Analysis: Comparison of Individual and Combined Performance of VNIR, XRF, and LIBS Spectroscopies. *Agronomy* **2021**, *11*, 1028. [CrossRef]
43. Chang, C.-W.; Laird, D.A.; Mausbach, M.J.; Hurburgh, C.R., Jr. Near-Infrared Reflectance Spectroscopy-Principal Components Regression Analyses of Soil Properties. *Soil Sci. Soc. Am. J.* **2001**, *65*, 480–490. [CrossRef]
44. Li, S.; Ji, W.; Chen, S.; Peng, J.; Zhou, Y.; Shi, Z. Potential of VIS-NIR-SWIR Spectroscopy from the Chinese Soil Spectral Library for Assessment of Nitrogen Fertilization Rates in the Paddy-Rice Region, China. *Remote Sens.* **2015**, *7*, 7029–7043. [CrossRef]
45. Wijewardane, N.K.; Ge, Y.; Morgan, C.L.S. Moisture insensitive prediction of soil properties from Vis-NIR reflectance spectra based on external parameter orthogonalization. *Geoderma* **2016**, *267*, 92–101. [CrossRef]
46. Franceschini, M.; Demattê, J.; Kooistra, L.; Bartholomeus, H.; Rizzo, R.; Fongaro, C.; Molin, J. Effects of external factors on soil reflectance measured on-the-go and assessment of potential spectral correction through orthogonalisation and standardisation procedures. *Soil Tillage Res.* **2018**, *177*, 19–36. [CrossRef]
47. Benedet, L.; Faria, W.M.; Silva, S.H.G.; Mancini, M.; Demattê, J.A.M.; Guilherme, L.R.G.; Curi, N. Soil texture prediction using portable X-ray fluorescence spectrometry and visible near-infrared diffuse reflectance spectroscopy. *Geoderma* **2020**, *376*, 114553. [CrossRef]
48. Kopačková, V.; Ben-Dor, E.; Carmon, N.; Notesco, G. Modelling Diverse Soil Attributes with Visible to Longwave Infrared Spectroscopy Using PLSR Employed by an Automatic Modelling Engine. *Remote Sens.* **2017**, *9*, 134. [CrossRef]
49. Velliangiri, S.; Alagumuthukrishnan, S.; Joseph, S.I.T. A Review of Dimensionality Reduction Techniques for Efficient Computation. *Procedia Comput. Sci.* **2019**, *165*, 104–111. [CrossRef]