applied sciences

MDPI

Article

# Potential Model Overfitting in Predicting Soil Carbon Content by Visible and Near-Infrared Spectroscopy

**Lizardo Reyna [1,2,†], Francis Dube [3], Juan A. Barrera [1] and Erick Zagal [1,*]**

[1] Department of Soils and Natural Resources, Faculty of Agronomy, Universidad de Concepción, Vicente Méndez 595, Casilla 537, Chillán 3812120, Chile; lreyna@udec.cl or lreyna@utm.edu.ec (L.R.); jbarrera@udec.cl (J.A.B.)

[2] Doctoral Program in Agronomic Sciences, Faculty of Agronomy, Universidad de Concepción, Vicente Méndez 595, Casilla 537, Chillán 3812120, Chile

[3] Department of Silviculture, Faculty of Forest Sciences, Universidad de Concepción, Victoria 631, Casilla 160-C, Concepción 4030000, Chile; fdube@udec.cl

* Correspondence: ezagal@udec.cl; Tel.: +56-42-2208853

† Current address: Facultad de Ingeniería Agrícola, Universidad Técnica de Manabí, Casilla 82, Lodana, Manabí, Ecuador.

Academic Editor: Johannes Kiefer

Received: 13 June 2017; Accepted: 5 July 2017; Published: 8 July 2017

**Abstract:** Soil spectroscopy is known as a rapid and cost-effective method for predicting soil properties from spectral data. The objective of this work was to build a statistical model to predict soil carbon content from spectral data by partial least squares regression using a limited number of soil samples. Soil samples were collected from two soil orders (Andisol and Ultisol), where the dominant land cover is native *Nothofagus* forest. Total carbon was analyzed in the laboratory and samples were scanned using a spectroradiometer. We found evidence that the reflectance was influenced by soil carbon content, which is consistent with the literature. However, the reflectance was not useful for building an appropriate regression model. Thus, we report here intriguing results obtained in the calibration process that can be confusing and misinterpreted. For instance, using the Savitzky–Golay filter for pre-processing spectral data, we obtained $R^2 = 0.82$ and root-mean-squared error $(RMSE) = 0.61\%$ in model calibration. However, despite these values being comparable with those of other similar studies, in the cross-validation procedure, the data showed an unusual behavior that leads to the conclusion that the model overfits the data. This indicates that the model should not be used on unobserved data.

**Keywords:** chemometrics; SOC; spectral diffuse reflectance; partial least squares regression; cross-validation

## 1. Introduction

Soil total carbon (TC) is composed of organic (all organic components mainly derived from the decomposition of plants and animals; and including living organisms) and inorganic (non-living C, typically as carbonates) carbon forms. Due to the short-term cycle of soil organic carbon (SOC) and its key role for soil functions, the quantitative evaluation of SOC is essential for determining a suitable management practice to conserve or increase soil carbon stock [1–4]. Monitoring SOC over large areas or long periods of time requires analysis of substantial numbers of samples which can be labor-intensive and expensive. Under those circumstances, the soil spectroscopy technique is an effective method to predict SOC rapidly at minimal cost [5,6]. Soil spectroscopy uses the visible and near-infrared (VIS-NIR, 400–2500 nm) and mid-infrared ( 2500–25,000 nm) spectral reflectance to infer soil properties from a scanned sample [7]. This technique has been used mainly under laboratory conditions, but it can also be applied in the field for a specific site or in an instrument setup for ongoing scanning [6].

Spectral reflectance of soil in the VIS-NIR has been used to predict soil C in different soil types. Sarkhot et al. [8] reported high correlation values for total and organic C ($R^2$ = 0.85 and $R^2$ = 0.86, respectively) with an error of 5.33 g·kg$^{-1}$ for total C and 2.88 g·kg$^{-1}$ for organic C in an entisol in the first 50-cm layer of soil. Fontán et al. [9] found higher correlation values for inorganic C ($R^2$ = 0.76) than organic C ($R^2$ = 0.67) in a vertisol to a depth of 90 cm. In addition, successful predictions of soil carbon fractions have been made by VIS-NIR spectroscopy [10,11], suggesting this technique as a reliable alternative for assessing the impact of land use change on soil carbon pools. Other related studies show the performance of different multivariate methods in calibrating models [12–14]. Classical methods for analyzing soil C require a rigorous sample preparation and expensive chemical supplies (e.g., dry combustion). As a complement to these methods, soil spectroscopy can be used to build empirical models to predict soil properties from spectral data. The potential of this technique to predict soil C has been extensively reported [15–19], and important considerations for the effective application of soil spectroscopy have been also reviewed by Reeves III [7]. However, to our knowledge, there are no studies that describe potential issues such as overfitting in model calibration procedures in a realistic field example.

In general, better correlations and fewer errors have been reported for sieved (<2 mm) and dry (air/oven) samples [9,20]. Brunet et al. [16] concluded that grinded samples notably improve the predictions of total soil carbon. In contrast, Fystro [15] found that the quality of prediction was not benefited by grinding samples. Soil darkness is the most evident effect of water on spectral variability, but strong effects occur in the infrared spectral reflectance where the wavelengths of 1450 nm and 1950 nm are absorption bands [21,22]. Because water molecules in soil are dispersed, broad bands are usually seen at these wavelengths [21]. Andisols have special properties such as low density and relatively high amounts of organic carbon within the soil profile [23], which have an important effect on soil reflectance.

Before model calibration, a pre-processing step is needed to reduce random noise and dimensionality of the spectral data [12]. The purpose of pre-processing techniques in spectroscopy is to eliminate the portion of reflectance that does not come from the desired properties of the target. In soil spectroscopy, the spectra variability includes incident light reflected in multiple directions (diffuse reflectance) due to soil roughness, soil aggregates, soil structure, and particle size. These physical properties scatter the incident energy in many directions. In a complex soil sample, the variation of reflectance is not strongly wavelength-dependent (Lorentz–Mie scattering) and can be observed as a multiplicative effect [24]. For this reason, filtering data is an essential process before analysis [25,26]. Spectral correction and spectral derivatives are two categories of pre-processing techniques. For the first category, the most used methods are multiplicative scatter correction (MSC) and the standard normal variate (SNV). For the second category, Norris–Williams (NW) and Savitzky–Golay (SG) derivatives filters are widely applied to spectral data. In this study, we performed the pre-processing step using the SG filter.

The aim of the present work was to evaluate the potential of soil NIR spectroscopy to predict total organic soil carbon using a limited number of soil samples collected from the native temperate forest. We used the partial least square regression (PLSR) method to find the statistical relation between the analyzed samples and spectral data.

## 2. Materials and Methods

### *2.1. Site Description*

The study took place in the Biobio and Araucanía Regions of Chile (Figure 1). Two soil orders were selected for the experiment: (1) Andisol, located in the Andean range (medial, amorphic, mesic, Typic Hapludands) [23]; and (2) Ultisol, which is located in the coastal range (very fine, mixed, semiactive, mesic Typic Paleudults) [27]. In both experimental sites the land cover is dominated by *Nothofagus obliqua* and *Nothofagus nervosa*. Andisol is located at latitude 36°48′ S, longitude

71°38′ W, and Ultisol is located at latitude 37°46′ S, longitude 72°58′ W. These soils are derived from volcanic ash due to the intense volcanic activity in the Quaternary [28] and are rich in organic carbon [23]. This parent material does not contain carbonates. Furthermore, in the field, these soils do not react to the hydrochloric acid test. Therefore, TC and organic carbon (OC) were equivalents.

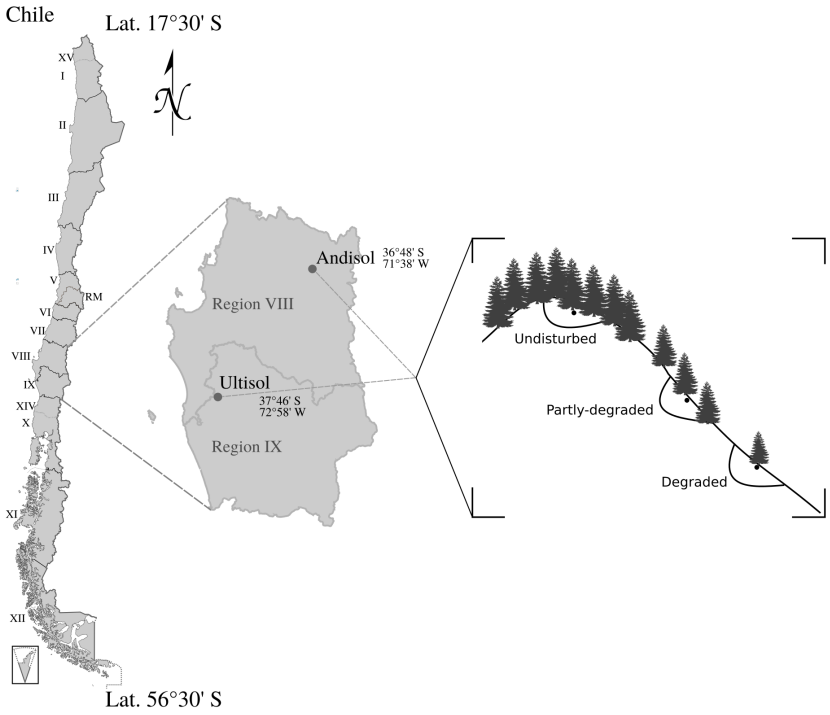


**Figure 1.** Location of the study sites. Two regions were selected; Region VIII (Biobio) and Region IX (Araucanía) which are located in the Andean and coastal ranges of Chile, respectively. In the box, a representation of the different *Nothofagus obliqua* forest conditions is shown for where the soil samples were extracted.

### *2.2. Soil Sampling, Total Carbon Analysis and Spectral Measurement*

A total of 70 samples were taken from the sites and three soil layers were sampled at depths of 0–5, 5–20, and 20–40 cm. The sampling strategy was selected to represent different conditions of the native forest, from undisturbed to degraded (Figure 1, for details see [27]). The samples were kept in plastic bags and stored at −4 °C for later analysis. Later, the samples were air-dried and put through a 2-mm sieve in order to minimize spectral variation due to fresh organic matter, soil moisture and soil aggregates. One portion of the sample was reserved for analysis and the other for spectral scanning. Total carbon (TC) was obtained by dry combustion method [29] using an elemental analyzer (model Leco CN-2000, macro-analyzer, LECO Corporation, Saint Joseph, Michigan, USA). All TC values were combined in a unique data set that contained the entire range of all depths and both soil types. Samples for scanning were oven-dried at 60 °C for 48 h to standardize the moisture level, then were set in petri dishes (60 × 15 mm) and flattened with a spatula. Thirty-six Andisol samples were taken and 34 Ultisol. Each depth was represented by twelve samples, except for the depths of 5–20 and 20–40 cm for Ultisol, where 11 samples were selected for each one.

Spectral reflectance was measured in the VIS-NIR range (350–1075 nm) at 1-nm intervals using a spectroradiometer (HandHeld 2: Hand-held VNIR, ASD-FieldSpec®, ASD, Boulder, Colorado, USA). We used this interval to detect possible sharp peaks in the spectral. The sensor was located vertically at a distance of 5 cm from the soil sample and it was fixed in a tripod, then a Spectralon panel was used as white reference before sample scans. Outdoor scanning was performed using the natural

source of light at 3:00 p.m. on a sunny day. A dark background was used to minimize the influence of ambient light, and to record the soil reflectance provided by natural illumination. Ten consecutive scans were averaged and recorded for each sample.

### 2.3. Spectral Pre-Processing and Reflectance Analysis

Before model calibration we performed several pre-processing configurations using the Savitzky–Golay filter in order to remove both additive and multiplicative effects in the spectral data [24]. This digital filter smooths the data while the original characteristics are minimally affected [30,31]. Input of three parameters are needed: window size, polynomial order, and derivative order e.g., (5, 1, 1). Window size must be set in the form 2M + 1, where M is half of the window size [30]. This value is the number of consecutive points that will be used by least-squares smoothing, and for the end points of the curve this window can be asymmetric. If derivative order is given, the data is transformed. In the original publication, a detailed explanation of this filter can be found [25]. We also evaluated two additional functions, the Log(1/R) (where R is reflectance) to transform reflectance (R) into apparent absorbance [5], and mean-center function (centering) making the spectral curves to fluctuate around zero [32]. We carried out fourteen pretreatments to the spectral data including smooth and transformation by first and second derivative orders in the Savitzky–Golay algorithm.

Spectral data can be decomposed into a small number of explanatory variables containing huge amount of variance [33]. Thus a principal component analysis (PCA) was performed on the spectral data to examine its structure and identify outliers.

The relationship between soil carbon content and reflectance is essentially inverse, i.e., when soil carbon increases the reflectance decreases across the spectra [19,34,35]. To prove this postulation, we conducted an analysis for each soil type in order to observe the influence of soil carbon content on reflectance. Thus, the spectral curves corresponding to each sampled depth were averaged and plotted for visual analysis. In the same way, TC values were also averaged to be related with the spectral reflectance (Table 1).

**Table 1.** Average of soil total carbon (TC) of each depth sampled for the two soil orders (Andisol and Ultisol). In brackets the numbers of samples averaged.

| Soil order | Depth (cm) | Average | Standard Deviation |
|---|---|---|---|
| Andisol | 0–5 | 6.4 (12) | 2.42 |
| | 5–20 | 4.4 (12) | 1.65 |
| | 20–40 | 3.0 (12) | 1.29 |
| Ultisol | 0–5 | 5.3 (12) | 1.55 |
| | 5–20 | 4.5 (11) | 1.26 |
| | 20–40 | 2.7 (11) | 1.78 |

### 2.4. Cross-Validation and Partial Least Squares Regression

Partial least squares regression or PLSR is a multivariate technique based on the combination of dimensionality reduction similar to PCA and multiple linear regression (MLR) [36]. Unlike PLSR, the latent variables or factors (equivalent to principal components in PCA) are calculated taking into account the response variable [33]. With these latent variables a predictor matrix **X** is built. Then, this matrix is used to build a regression model to explain the variance of the response variable **Y**. When **Y** contains only one variable, the method is referred to as PLSR1, and when it contains more than one variable it is referred to as PLSR2; in both cases we refer to Y as a matrix of response variables. Some exhaustive reviews of this method have been published [36,37]. PLSR1 was used to relate TC with spectral data.

A common practice in soil spectroscopy is to use cross-validation techniques to select the optimal number of latent variables for PLSR. We performed leave-*k*-out cross-validation [38,39] to select the

number of latent variables for regression (also referred to as $k$-fold CV). In this technique, the data set is randomly split into $k$ equal-sized groups, where $k$ is defined by the user. We used $k = 5$ as recommended by Li et al. [39] and latent variables ranged from 1 to 20. One group is left out, and the model is calibrated with the remaining $(k - 1)$ groups. Then, the prediction accuracy of the model is evaluated using the group that was left out by comparing the predicted and measured values. This process is repeated until each group is used to validate the model. When the value of $k$ is equal to the number of samples, this technique is referred to as leave-one-out cross-validation (LOOCV). The mean of statistical indicators such as root-mean-squared error (*RMSE*) (Equation (1)) and coefficient of determination $R^2$ (Equation (2)) were used as a reference to evaluate the model performance.

$$RMSE = \sqrt{\sum_{i=1}^{i} \frac{(y_i - \hat{y}_i)^2}{n}} \quad (1)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{i}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{i}(y_i - \overline{y}_i)^2} \quad (2)$$

where $y_i$ is the measured value of sample $i$, $\hat{y}_i$ is the predicted value of sample $i$ and $\overline{y}$ is the average of measured samples.

After the optimal number of latent variables were determined, the data set was split into two groups. The group for calibration contained 80% of the data set and the remaining 20% was used as validation group. Some strategies to split data have been applied by researchers [18,33,40]. Nonetheless, we performed an iterative procedure to find the best data split (80/20%) in terms of model performance i.e., several possible data splits of 80/20% were evaluated by PLSR. This was made using the scikit-learn library [41] of Python programming language (Python Software Foundation, https://www.python.org/), which achieves the same data split every time (pseudo-random) in order to reproduce the results. The best calibration/validation sets were selected by the higher $R^2$ in PLSR.

For the final model selection, all pre-processing configurations listed in Table 2 were applied to the spectral data before splitting. Then, PLSR was performed for each one and the quality of model prediction was evaluated using the most-used quality estimators for regression such as $R^2$ and *RMSE*. All analyses were performed using a Python-based ecosystem for scientific computing [42] (https://www.scipy.org/). All calculations ware made under the Jupyter Notebook environment (formerly Ipython Notebook) to facilitate the reproduction of the results [43] (Supplementary Files: S1-reflectance-analysis and S2-calculations).

## 3. Results and Discussion

### *3.1. Soil Total Carbon*

Total carbon measured in the 70 samples ranged from 0.77 to 10.7% (Figure 2A). The highest and lowest concentration of C in soil were found in the Andisol at the depths of 0–5 and 20–40 cm, respectively. These values were justified by environmental conditions and soil type. For Andisol, the ranges of TC of the three depths overlap mainly in lower values, i.e., values between 2 and 5% were found in all of the sampled depths. Values greater than 8% were found in the top soil layer (0–5 cm). For Ultisol, the lower values of TC were clearly observed in the depth of 20–40 cm. However, outliers were found for values greater than 4%, probably due to such samples being taken from undisturbed *Nothofagus* forest conditions [27]. The minimum and maximum frequency of the TC values were 1 and 14 (Figure 2B), respectively.

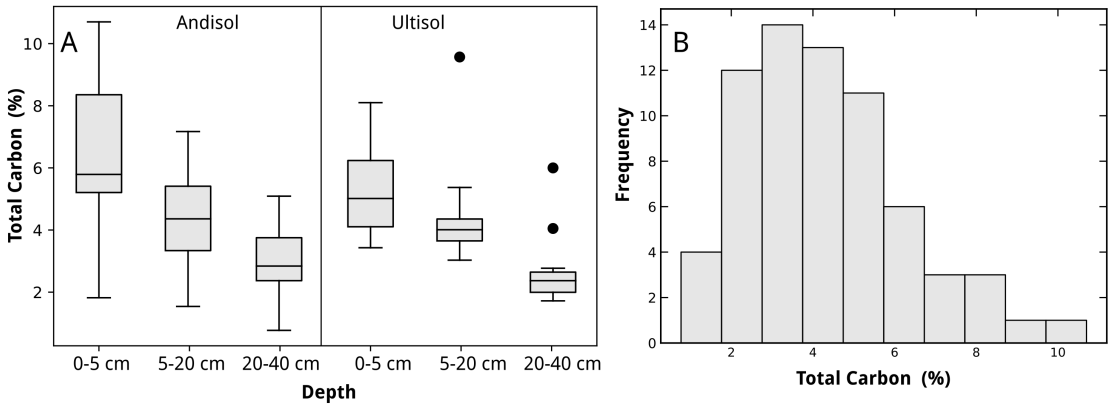


**Figure 2.** Carbon content by depth and soil order. (**A**) The full range variation is represented in the box and whisker diagrams. The top of the gray box is the third quartile and the bottom the first quartile. The horizontal line inside the box is the median of the data. Whiskers above and below the box show the minimum and maximum values and outliers are represented by black circles; (**B**) The distribution of the soil total carbon of the whole data set.

### *3.2. Spectral Pre-Processing*

Before filtering the entire spectral data set, we tested the Savitzky–Golay filter on a noisy spectral curve in order to visually inspect its effect on noise (Figure 3). These filter settings (no transformation) were applied to the entire spectral data and their means were compared; their means showed no significant differences. The extremes of the curves were removed due to noise, and the range of 400–924 nm was selected for the analysis.

After pre-processing, PCA was performed on the filtered data sets to explore their structure. In the most cases, two principal components explained high amount of variability of the spectra (>95%). However, using first derivative transformation, the explained variability reached 70% with five PCs and this value decreased when derivative order increased. Nonetheless, PCA showed potential for classifying samples from different soil orders from spectral data. A sophisticated method for enhancing the performance of classification of soils from spectral data was proposed by Xie et al. [44].

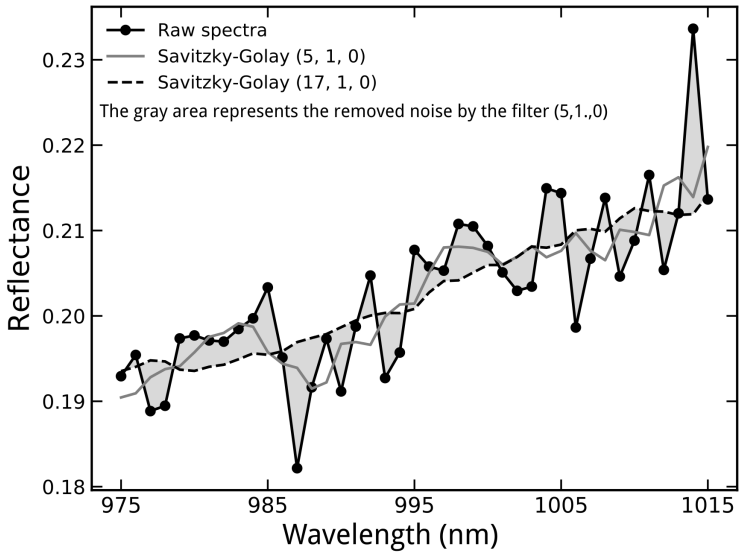


**Figure 3.** Effect of the Savitzky–Golay filter applied to a noisy spectral curve. No derivative transformation was used. Only forty points are displayed in order to appreciate its effect on noise. Inside the parentheses, window size, polynomial order and derivative order are given.

### *3.3. Effect of Soil TC on Reflectance*

As we expected, soil layers with higher TC values, tend to have lower reflectance. However, in Andisol samples (Figure 4A), for 5–20 and 20–40 cm depths, the spectral curves intersected at wavelengths shorter than 400 nm. By contrast, the maximum spectral separation was observed near 650 nm. In the Ultisol samples (Figure 4B), the curves of the depths of 0–5 and 5–20 cm intersected between 700 and 800 nm. We attribute this curve intersection to the little difference in soil TC between the first two depths in Ultisol soil (Figure 5). We also calculated the correlation coefficient ($r$) between every spectral curve and soil TC. In accordance with [19], we found the most negative correlation between TC and spectral wavelengths near the 500–600 nm range (Figure 6).

By comparing soil reflectance with TC, our results showed the influence of absorption features of soil carbon on reflectance. This is in agreement with previous works published by [19,34]. However, the correlation coefficient between spectral bands and TC in the region of 500–700 nm ($r \approx -0.5$), were poorer than those reported by others [19,34] ($r = -0.8$ or better). One possible explanation of this is the spectral distortion due to natural source of light used for scanning the samples (outdoor scanning) and low bulk density of the studied soils ($\leq 1$ g$\cdot$ cm$^{-3}$) [27] which promote more dispersion of light [45]. This distortion was probably not successfully corrected by the Savitzky–Golay filter. Despite this, near 450 nm (Figure 4), the absorption features of distinct iron oxides can be observed [34]. For the Ultisol samples, this absorption was lightly observed in the curve. The spectral separability was not clear when the curves were plotted individually, for some samples with higher soil TC the reflectance along the spectra was not necessarily lower than for samples with lower TC. However, our results showed that averaging the reflectance by depth, the influence of soil TC was evident.

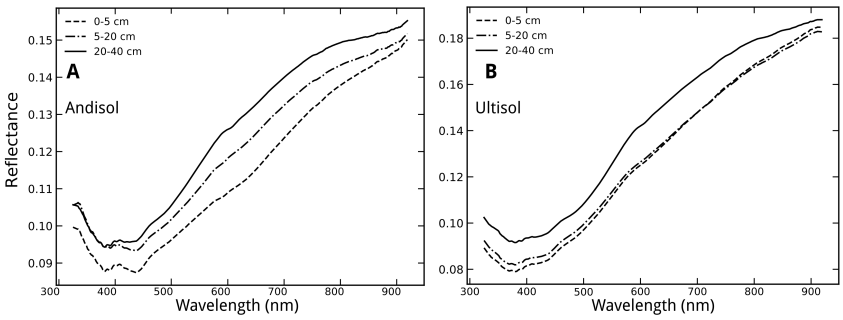


**Figure 4.** The average of soil reflectance of the three soil depths. (**A**) The spectral separability between depths can be observed from 450 nm; (**B**) for the depths of 0–5 and 5–20 cm the spectral separability was not exhibited, and the curves intersected between 650 and 750 nm.

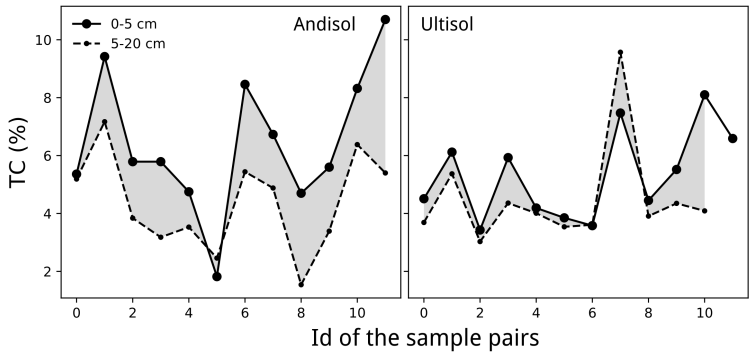


**Figure 5.** Total carbon (TC) in the Andisol and Ultisol soil orders. The gray area represents the difference in soil TC of the two depths.

### *3.4. Cross-Validation (CV) and Partial Least Squares Regression*

To better understand the experimental results in this section, we will discuss them separately. Firstly, we will discuss the results obtained in leave-*k*-out CV, secondly we will analyze the results obtained by splitting the data set into calibration/validation subsets (80/20%, respectively), and finally the implications of the potential misinterpretation of the results will be addressed.

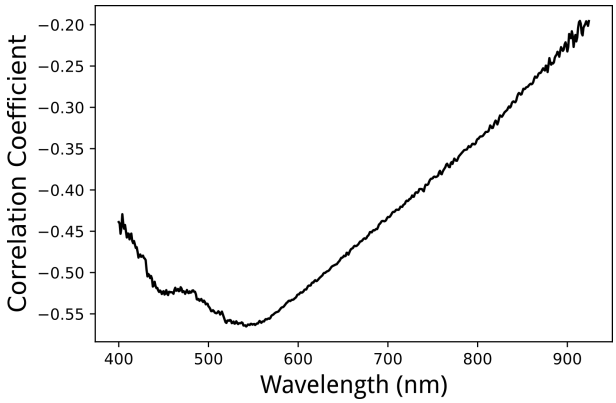


**Figure 6.** Correlation coefficient between individual spectral curves and TC. This graph shows the negative correlation between soil carbon content and soil reflectance. The higher correlation occurs near 550 nm.

#### 3.4.1. Cross-Validation

With the parameters used for CV, the smoothed spectra yielded the lowest *RMSE* (1.8%) with 2 latent variables (LVs). The performance of the model was significantly poorer from 6 LVs (Figure 7) and this did not improve using LOOCV which is considered the most accurate [39]. If this is the case, Abdi [46] manifests that the model is overfitting the data and is not useful for predicting unobserved data. To our knowledge, this unusual behavior of the data in cross-validation has rarely been reported in the literature. According to several authors [39,47,48] the optimal number of LVs is determined when only marginal improvements in model performance are observed, but with our data this marginal improvement was not clear, and the optimal number of LVs was selected based on lowest *RMSE* (Figure 7A). The $R^2$ in most cases was negative, indicating a bad model fit (Figure 7B). Negative $R^2$ resulted in model calibration, which has been also reported [19].

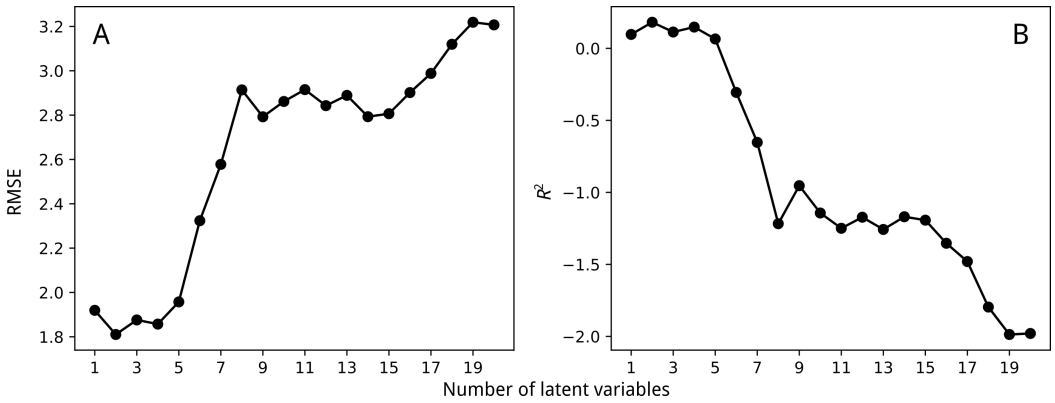


**Figure 7.** Performance of partial least squares regression (PLSR) in cross-validation with $k = 5$. (**A**) Root-mean-squared error (*RMSE*), and (**B**) $R^2$. These graphs could be used to indicate the consistency of the data for modeling. Both curves indicate that the data is not useful for modeling.

### 3.4.2. PLSR Calibration

Prior to final model calibration, an iterative process was performed to evaluate 600 possible data splits into calibration and validation data sets, at 80/20%, respectively. The best data split was selected by the highest $R^2$ resulting from PLSR prediction. The key function for this process was train_test_split of the scikit-learn library, test_size parameter was 0.2 indicating the proportion of the validation set. The optimal value of random_state parameter was 280 in most cases.

Using the split data set of the smoothed spectra, the performance of the model was significantly better in terms of $R^2$ and *RMSE*. Varying the number of LVs to 1 and 5, the values of $R^2$ were 0.82 and 0.74 and *RMSE* were 0.64% and 0.75%, respectively. Usingmore than 5 LVs, the $R^2$ decreased drastically. With some pre-processing settings the model fitted better than others (Table 2). This has been also found by several authors [16,19,26]. An ordinary least squares regression was performed to inspect the best model (Figure 8). Our best model ($R^2$ = 0.82 and *RMSE* = 0.64%) indicates a good prediction capacity for TC in accordance with the standard used by Sarkhot et al. [8], and the worst model was obtained using a second-order derivative transformation ($R^2$ = 0.23 and RSME = 1.51%). These results were congruent with those reported by [49] who used a similar pre-processing configuration, and the second derivative transformation produced the worst models for the most of the predicted variables. In contrast, Vasques et al. [12] found subtle differences in PLSR performance using derivative transformations in pre-processing the data. On the other hand, Knadel et al. [50] indicated that non-preprocessed spectral data generated the best model for soil organic carbon. Window size and polynomial order in the Savitzky–Golay algorithm had no important influence on model performance (Table 2).

Our results demonstrate that with the same data, it was possible to obtain different results in predicting soil TC. This discordance between the two procedures (with and without cross-validation), may lead to an ambiguous interpretation of the predictive capacity of the built model. For example, avoiding cross-validation, the model resulted in a good capacity for predicting soil TC and could be used to predict TC from soil samples. Before final model calibration, the cross-validation technique was useful to inspect the potential of the spectral data to predict soil TC, and to select the optimal number of the latent variables for regression. PLSR is subject to overprediction when the number of samples for calibration is small (<58, in the soil spectroscopy context) [22]. However, some studies applied soil spectroscopy with fewer samples [12]. We used 70 soil samples from two study sites. After splitting the data and removing outliers, the calibration set resulted with 54 samples in most cases. The Ultisol samples were re-scanned several days later and the results were similar despite a different intensity of reflectance. More research is needed to better understand soil reflectance variability in these soil types.

**Table 2.** PLSR performance for different Savitzky–Golay filter configurations applied to spectral data. The outlier ID column contains the index of the samples in the data set that were considered as outliers.

| SG Filter | Number of LVs | $R^2$ | *RMSE* | Outliers ID |
|---|---|---|---|---|
| (5, 1, 0) | 2 | 0.82 | 0.61 | 21, 60 |
| (5, 2, 0) | 2 | 0.82 | 0.61 | 21, 60 |
| (5, 1, 1) | 2 | 0.58 | 1.22 | 20, 57 |
| (5, 2, 2) | 1 | 0.23 | 1.51 | – |
| (11, 1, 0) | 2 | 0.82 | 0.61 | 21, 60 |
| (11, 2, 0) | 2 | 0.82 | 0.61 | 21, 60 |
| (11, 1, 1) | 1 | 0.54 | 1.08 | – |
| (11, 2, 2) | 1 | 0.36 | 1.29 | 20 |
| (17, 1, 0) | 2 | 0.82 | 0.61 | 21, 60 |
| (17, 2, 0) | 2 | 0.82 | 0.61 | 21, 60 |
| (17, 1, 1) | 2 | 0.58 | 1.48 | – |
| (17, 2, 2) | 1 | 0.26 | 1.59 | 20 |
| (17, 1, 0) + Log(1/R) | 2 | 0.79 | 0.66 | 21, 60 |
| (17, 1, 0) + centering | 2 | 0.79 | 0.66 | 21, 60 |

In summary we found that firstly, the average of reflectance in the VIS-NIR recorded from soil samples was useful as a descriptive type of information about the carbon content, showing an appreciable relationship with soil TC, especially in the andisol samples. However, it was not appropriate for building a robust model for predicting soil TC. Secondly, the Savitzky–Golay filter was effective in eliminating the most visible noise in the spectral data. Thirdly, using the random_state parameter in the iterative process, we rapidly found the best calibration/validation subsets for model calibration.

We attempt with this study to offer additional support for an effective application of soil spectroscopy. In contrast to most published works, we report this negative case of the soil spectroscopy technique to show the potential model overfitting and misinterpretation of the results by soil scientists with little experience in data analysis. The results presented here need to be interpreted with caution because of the unusual behavior observed in reflectance for predicting soil carbon content.

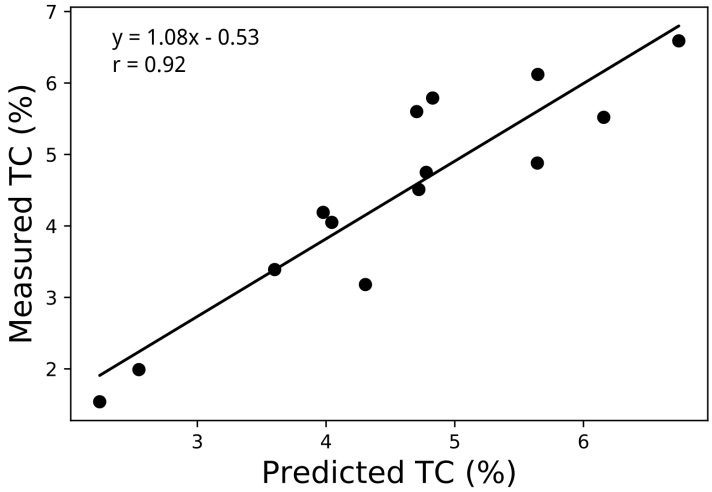


**Figure 8.** Linear regression applied to predicted and measured TC.

## 4. Conclusions

We conducted the standard procedure to build a statistical model to predict soil TC from spectral data. The results of the experiment warn of possible model overfitting when the sources of variability of the spectra (particle size and illumination) have not been effectively controlled and the amount and distribution of the soil samples are inadequate. However, we demonstrate that if cross-validation (CV) is avoided, it is possible to obtain a good PLSR model, which may, in turn, be inappropriately applied to unobserved data. To identify this issue, the cross-validation technique was useful for plotting the performance of the model versus the number of latent variables. Notwithstanding, our results confirm that the reflectance was influenced by soil carbon content, although it was only useful at the description level. We concluded that the potential of soil spectroscopy may be minimized when the spectral distortion exceeds the capacity of the filter to correct it. The cause of the spectral variability in these soil samples remains unclear. More sophisticated instruments and more rigorous scanning procedures may help to understand why in this case soil spectroscopy was not feasible.

**Supplementary Materials:** The following are available online at www.mdpi.com/2072-6651/7/7/708/s1, S1: reflectance-analysis, S2: calculations.

**Acknowledgments:** The present work benefited from the FONDECYT Project No. 11121279 granted by the National Commission for Scientific and Technological Research of the Chilean Government. L.R. received financial support from the National Secretary of Higher Education, Science, and Technology (SENESCYT) of Ecuador. Two anonymous reviewers are acknowledged for their suggestions that improved the manuscript.

**Author Contributions:** Lizardo Reyna, Juan A. Barrera and Erick Zagal, conceived, designed and performed the experiments; Francis Dube facilitated access to the experimental sites and soil samples. All authors contributed with valuable discussions and scientific advices in order to improve the quality of the work.

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

## References


1. Stevenson, F.; Cole, M. *Cycles of Soils*, 2nd ed.; Wiley: New York, NY, USA, 1999.
2. Dube, F.; Zagal, E.; Stolpe, N.; Espinosa, M. The Influence of Land-Use Change on the Organic Carbon Distribution and Microbial Respiration in a Volcanic Soil of the Chilean Patagonia. *For. Ecol. Manag.* **2009**, *257*, 1695–1704.
3. Zagal, E.; Muñoz, C.; Espinoza, S.; Campos, J. Soil Profile Distribution of Total C Content and Natural Abundance of 13C in Two Volcanic Soils Subjected to Crop Residue Burning versus Crop Residue Retention. *Acta Agric. Scand.* **2012**, *62*, 263–272.
4. Powlson, D.; Smith, P.; Nobili, M.D. Soil organic matter. In *Soil Conditions and Plant Growth*; Blackwell Publishing Ltd.: Oxford, UK, 2013; pp. 86–131.
5. Stenberg, B.; Viscarra Rossel, R.A.; Mouazen, A.M.; Wetterlind, J. Chapter Five—Visible and Near Infrared Spectroscopy in Soil Science. In *Advances in Agronomy*; Sparks, D.L., Ed.; Academic Press: San Diego, CA, USA, 2010; Volume 107, pp. 163–215.
6. Viscarra Rossel, R.A.; Adamchuk, V.I.; Sudduth, K.A.; McKenzie, N.J.; Lobsey, C. Chapter Five—Proximal Soil Sensing: An Effective Approach for Soil Measurements in Space and Time. In *Advances in Agronomy*; Sparks, D.L., Ed.; Academic Press: San Diego, CA, USA, 2011; Volume 113, pp. 243–291.
7. Reeves, J.B., III. Near- versus Mid-Infrared Diffuse Reflectance Spectroscopy for Soil Analysis Emphasizing Carbon and Laboratory versus on-Site Analysis: Where Are We and What Needs to Be Done? *Geoderma* **2010**, *158*, 3–14.
8. Sarkhot, D.V.; Grunwald, S.; Ge, Y.; Morgan, C.L.S. Comparison and Detection of Total and Available Soil Carbon Fractions Using Visible/near Infrared Diffuse Reflectance Spectroscopy. *Geoderma* **2011**, *164*, 22–32.
9. Fontán, J.M.; Calvache, S.; López-Bellido, R.J.; López-Bellido, L. Soil Carbon Measurement in Clods and Sieved Samples in a Mediterranean Vertisol by Visible and Near-Infrared Reflectance Spectroscopy. *Geoderma* **2010**, *156*, 93–98.
10. Reeves, J.B., III; Follett, R.F.; McCarty, G.W.; Kimble, J.M. Can Near or Mid-Infrared Diffuse Reflectance Spectroscopy Be Used to Determine Soil Carbon Pools? *Commun. Soil Sci. Plant Anal.* **2006**, *37*, 2307–2325.
11. Knox, N.M.; Grunwald, S.; McDowell, M.L.; Bruland, G.L.; Myers, D.B.; Harris, W.G. Modelling Soil Carbon Fractions with Visible Near-Infrared (VNIR) and Mid-Infrared (MIR) Spectroscopy. *Geoderma* **2015**, *239*, 229–239.
12. Vasques, G.M.; Grunwald, S.; Sickman, J.O. Comparison of Multivariate Methods for Inferential Modeling of Soil Carbon Using Visible/near-Infrared Spectra. *Geoderma* **2008**, *146*, 14–25.
13. Lucà, F.; Conforti, M.; Castrignanò, A.; Matteucci, G.; Buttafuoco, G. Effect of Calibration Set Size on Prediction at Local Scale of Soil Carbon by Vis-NIR Spectroscopy. *Geoderma* **2017**, *288*, 175–183.
14. Mouazen, A.M.; Kuang, B.; De Baerdemaeker, J.; Ramon, H. Comparison among Principal Component, Partial Least Squares and Back Propagation Neural Network Analyses for Accuracy of Measurement of Selected Soil Properties with Visible and near Infrared Spectroscopy. *Geoderma* **2010**, *158*, 23–31.
15. Fystro, G. The Prediction of C and N Content and Their Potential Mineralisation in Heterogeneous Soil Samples Using Vis-NIR Spectroscopy and Comparative Methods. *Plant Soil* **2002**, *246*, 139–149.
16. Brunet, D.; Barthès, B.G.; Chotte, J.L.; Feller, C. Determination of Carbon and Nitrogen Contents in Alfisols, Oxisols and Ultisols from Africa and Brazil Using NIRS Analysis: Effects of Sample Grinding and Set Heterogeneity. *Geoderma* **2007**, *139*, 106–117.
17. Gomez, C.; Viscarra Rossel, R.A.; McBratney, A.B. Soil Organic Carbon Prediction by Hyperspectral Remote Sensing and Field Vis-NIR Spectroscopy: An Australian Case Study. *Geoderma* **2008**, *146*, 403–411.
18. Wenjun, J.; Zhou, S.; Jingyi, H.; Shuo, L. In Situ Measurement of Some Soil Properties in Paddy Soil Using Visible and Near-Infrared Spectroscopy. *PLoS ONE* **2014**, *9*, e105708.
19. Zheng, G.; Ryu, D.; Jiao, C.; Hong, C. Estimation of Organic Matter Content in Coastal Soil Using Reflectance Spectroscopy. *Pedosphere* **2016**, *26*, 130–136.

20. Guillén, C.E.; Dávila, M.J.; Gilliot, J.M.; Vaoudour, E. Aporte de la espectroscopia a la estimación de carbono orgánico de los suelos de la planicie de Versalles, Francia. *Revista Geográfica Venezolana* **2013**, *54*, 85–98.
21. Baumgardner, M.F.; Silva, L.F.; Biehl, L.L.; Stoner, E.R. Reflectance Properties of Soils. In *Advances in Agronomy*; Brady, N.C., Ed.; Academic Press: San Diego, CA, USA, 1986; Volume 38, pp. 1–44.
22. Reeves, J.B., III; McCarty, G.W.; Calderon, F.; Hively, W.D. Chapter 20—Advances in Spectroscopic Methods for Quantifying Soil Carbon A2—Liebig, Mark A. In *Managing Agricultural Greenhouse Gases*; Franzluebbers, A.J., Follett, R.F., Eds.; Academic Press: San Diego, CA, USA, 2012; pp. 345–366.
23. Stolpe, N.B. *Descripción de Los Principales Suelos de La VII Región de Chile*; Publicaciones del Departamento de Suelos y Recursos Naturales—Universidad de Concepción: Chillán, Chile, 2006; Volume 1, p. 1.
24. Rinnan, Å.; van den Berg, F.; Engelsen, S.B. Review of the Most Common Pre-Processing Techniques for near-Infrared Spectra. *TrAC Trends Anal. Chem.* **2009**, *28*, 1201–1222.
25. Savitzky, A.; Golay, M.J.E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.* **1964**, *36*, 1627–1639.
26. Nawar, S.; Buddenbaum, H.; Hill, J.; Kozak, J.; Mouazen, A.M. Estimating the Soil Clay Content and Organic Matter by Means of Different Calibration Methods of Vis-NIR Diffuse Reflectance Spectroscopy. *Soil Tillage Res.* **2016**, *155*, 510–522.
27. Dube, F.; Stolpe, N.B. SOM and Biomass C Stocks in Degraded and Undisturbed Andean and Coastal Nothofagus Forests of Southwestern South America. *Forests* **2016**, *7*, 320.
28. Casanova, M.; Salazar, O.; Seguel, O.; Luzio, W. Main Features of Chilean Soils. In *The Soils of Chile*; Springer: Dordrecht, The Netherlands, 2013; pp. 25–97.
29. Wright, A.F.; Bailey, J.S. Organic Carbon, Total Carbon, and Total Nitrogen Determinations in Soils of Variable Calcium Carbonate Contents Using a Leco CN-2000 Dry Combustion Analyzer. *Commun. Soil Sci. Plant Anal.* **2001**, *32*, 3243–3258.
30. Schafer, R.W. What Is a Savitzky-Golay Filter? [Lecture Notes]. *IEEE Signal Process. Mag.* **2011**, *28*, 111–117.
31. Kinoshita, R.; Roupsard, O.; Chevallier, T.; Albrecht, A.; Taugourdeau, S.; Ahmed, Z.; van Es, H.M. Large Topsoil Organic Carbon Variability Is Controlled by Andisol Properties and Effectively Assessed by VNIR Spectroscopy in a Coffee Agroforestry System of Costa Rica. *Geoderma* **2016**, *262*, 254–265.
32. Van den Berg, R.A.; Hoefsloot, H.C.; Westerhuis, J.A.; Smilde, A.K.; van der Werf, M.J. Centering, Scaling, and Transformations: Improving the Biological Information Content of Metabolomics Data. *BMC Genom.* **2006**, *7*, 142.
33. Adeline, K.R.M.; Gomez, C.; Gorretta, N.; Roger, J.M. Predictive Ability of Soil Properties to Spectral Degradation from Laboratory Vis-NIR Spectroscopy Data. *Geoderma* **2017**, *288*, 143–153.
34. Henderson, T.L.; Baumgardner, M.F.; Franzmeier, D.P.; Stott, D.E.; Coster, D.C. High Dimensional Reflectance Analysis of Soil Organic Matter. *Soil Sci. Soc. Am. J.* **1992**, *53*, 865–872.
35. Zhang, P.; Shao, M. Spatial Variability and Stocks of Soil Organic Carbon in the Gobi Desert of Northwestern China. *PLoS ONE* **2014**, *9*, e93584.
36. Geladi, P.; Kowalski, B.R. Partial Least-Squares Regression: A Tutorial. *Anal. Chim. Acta* **1986**, *185*, 1–17.
37. Haenlein, M.; Kaplan, A.M. A Beginner's Guide to Partial Least Squares Analysis. *Underst. Stat.* **2004**, *3*, 283–297.
38. Jonathan, P.; Krzanowski, W.J.; McCarthy, W.V. On the Use of Cross-Validation to Assess Performance in Multivariate Prediction. *Stat. Comput.* **2000**, *10*, 209–229.
39. Li, B.; Morris, J.; Martin, E.B. Model Selection for Partial Least Squares Regression. *Chemom. Intell. Lab. Syst.* **2002**, *64*, 79–89.
40. Nocita, M.; Stevens, A.; Noon, C.; van Wesemael, B. Prediction of Soil Organic Carbon for Different Levels of Soil Moisture Using Vis-NIR Spectroscopy. *Geoderma* **2013**, *199*, 37–42.
41. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
42. Jones, E.; Oliphant, T.; Peterson, P. SciPy: Open Source Scientific Tools for Python. Available online: http://www.scipy.org (accessed on 5 May 2016).
43. Shen, H. Interactive Notebooks: Sharing the Code. *Nat. News* **2014**, *515*, 151.

44. Xie, H.; Zhao, J.; Wang, Q.; Sui, Y.; Wang, J.; Yang, X.; Zhang, X.; Liang, C. Soil Type Recognition as Improved by Genetic Algorithm-Based Variable Selection Using near Infrared Spectroscopy and Partial Least Squares Discriminant Analysis. *Sci. Rep.* **2015**, *5*, 10930.
45. Demattê, J.A.M.; Nanni, M.R.; da Silva, A.P.; de Melo Filho, J.F.; Santos, W.C.D.; Campos, R.C. Soil Density Evaluated by Spectral Reflectance as an Evidence of Compaction Effects. *Int. J. Remote Sens.* **2010**, *31*, 403–422.
46. Abdi, H. Partial Least Squares Regression and Projection on Latent Structure Regression (PLS Regression). *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 97–106.
47. Brown, D.J.; Bricklemyer, R.S.; Miller, 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.
48. Viscarra Rossel, R.A. ParLeS: Software for Chemometric Analysis of Spectroscopic Data. *Chemom. Intell. Lab. Syst.* **2008**, *90*, 72–83.
49. Askari, M.S.; O'Rourke, S.M.; Holden, N.M. Evaluation of Soil Quality for Agricultural Production Using Visible-near-Infrared Spectroscopy. *Geoderma* **2015**, *243–244*, 80–91.
50. Knadel, M.; Gislum, R.; Hermansen, C.; Peng, Y.; Moldrup, P.; de Jonge, L.W.; Greve, M.H. Comparing Predictive Ability of Laser-Induced Breakdown Spectroscopy to Visible near-Infrared Spectroscopy for Soil Property Determination. *Biosyst. Eng.* **2017**, *156*, 157–172.