




## Article

# Estimation of Organic Carbon in Anthropogenic Soil by VIS-NIR Spectroscopy: Effect of Variable Selection

Lu Xu <sup>1</sup> , Yongsheng Hong <sup>1,2</sup>, Yu Wei <sup>1</sup>, Long Guo <sup>3</sup>, Tiezhu Shi <sup>4</sup>, Yi Liu <sup>5</sup>, Qinghu Jiang <sup>6</sup>, Teng Fei <sup>1</sup> , Yaolin Liu <sup>1</sup>, Abdul M. Mouazen <sup>2</sup> and Yiyun Chen <sup>1,7,\*</sup> 

<sup>1</sup> School of Resource and Environmental Sciences, Wuhan University, Wuhan 430079, China; xuluwh@whu.edu.cn (L.X.); hys@whu.edu.cn (Y.H.); weixyu@whu.edu.cn (Y.W.); feiteng@whu.edu.cn (T.F.); yaolin610@whu.edu.cn (Y.L.)

<sup>2</sup> Department of Environment, Ghent University, Coupure Links 653, 9000 Gent, Belgium; abdul.mouazen@ugent.be

<sup>3</sup> College of Resources and Environment, Huazhong Agricultural University, Wuhan 430070, China; guolong@mail.hzau.edu.cn

<sup>4</sup> School of Architecture and Urban Planning & MNR Key Laboratory for Geo-Environmental Monitoring of Great Bay Area & Guangdong Key Laboratory of Urban Informatics & Shenzhen Key Laboratory of Spatial Smart Sensing and Services, Shenzhen University, Shenzhen 518060, China; tiezhushi@szu.edu.cn

<sup>5</sup> School of Public Administration, Guangdong University of Finance & Economics, Guangzhou 510320, China; liuyi@gdufe.edu.cn

<sup>6</sup> Key Laboratory of Aquatic Botany and Watershed Ecology, Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan 430074, China; jiangqh@wbcas.cn

<sup>7</sup> State Key Laboratory of Soil and Sustainable Agriculture, Chinese Academy of Sciences, Nanjing 210008, China

\* Correspondence: chenyy@whu.edu.cn; Tel.: +86-177-8648-6713

Received: 14 August 2020; Accepted: 13 October 2020; Published: 16 October 2020



**Abstract:** Visible and near-infrared reflectance (VIS-NIR) spectroscopy is widely applied to estimate soil organic carbon (SOC). Intense and diverse human activities increase the heterogeneity in the relationships between SOC and VIS-NIR spectra in anthropogenic soil. This fact results in poor performance of SOC estimation models. To improve model accuracy and parsimony, we investigated the performance of two variable selection algorithms, namely competitive adaptive reweighted sampling (CARS) and random frog (RF), coupled with five spectral pretreatments. A total of 108 samples were collected from Jiangnan Plain, China, with the SOC content and VIS-NIR spectra measured in the laboratory. Results showed that both CARS and RF coupled with partial least squares regression (PLSR) outperformed PLSR alone in terms of higher model accuracy and less spectral variables. It revealed that spectral variable selection could identify important spectral variables that account for the relationships between SOC and VIS-NIR spectra, thereby improving the accuracy and parsimony of PLSR models in anthropogenic soil. Our findings are of significant practical value to the SOC estimation in anthropogenic soil by VIS-NIR spectroscopy.

**Keywords:** anthropogenic soil; spectral variable selection; soil organic carbon; visible and near-infrared spectroscopy

## 1. Introduction

Soil holds the most massive storage of organic carbon in the terrestrial ecosystems [1,2]. Natural soil may transform into anthropogenic soil by long-term human activities [3]. Soil organic carbon (SOC) storage has also been altered due to this transformation [4]. Given that SOC plays a pivotal role in global warming [5,6], carbon cycle [7,8], and food security [6,9], changes of SOC storage may influence

the balance of ecosystem service. Therefore, monitoring SOC content in anthropogenic soil becomes a critical and urgent task [10,11].

Visible and near-infrared (VIS-NIR) spectroscopy serves as a time-saving and cost-effective proximal soil-sensing technique to estimate SOC content [12–15]. VIS-NIR spectra can be measured in situ or the lab [16]. However, VIS-NIR spectra measured in situ easily suffer from the influences of soil surface roughness, soil moisture, water vapor, light intensity, and other external environmental interference [17]. The measurement of VIS-NIR spectra performed in the lab could effectively avoid these influences.

Intensive human activities increase the heterogeneity in the relationship between SOC and VIS-NIR spectra, which leads to a new challenge to apply VIS-NIR spectroscopy for SOC estimation [18]. Several soil spectral libraries (SSLs) have been established to cover the heterogeneous spectral characteristics of different soil types as comprehensive as possible, thereby improving the accuracy of the SOC estimation. Existing SSLs include ICRAF-ISRIC world soil spectral library [19], European spectral library [20], Brazilian soil spectral library [21], Australian soil spectroscopic database [22], China soil spectral library [23], etc. These soil samples in the spectral libraries are collected on a large scale with low sampling density. For example, the mean soil sampling density of the European spectral library is 77 samples per 10,000 km<sup>2</sup> [24]. It is not enough to reveal the heterogeneous relationship between the VIS-NIR spectra and SOC on a small scale [25]. Therefore, many researchers made efforts on small-scale studies with soil samples collected from farmland, when VIS-NIR spectroscopy showed good performance [26–30]. These farmlands are continuous and have a large area with similar human activities. However, the heterogeneity in the relationship between SOC and VIS-NIR spectra is more complex in the highly fragmented farmland with various human activities, which can weaken the performance of the SOC estimation model by VIS-NIR spectra. To improve model performance, previous studies adopted the strategy of using representative calibration samples [26–29]. Nevertheless, the efficiency of this strategy is susceptible to sample size [30]. Therefore, it is essential to investigate other approaches that may improve model accuracy. Besides, VIS-NIR spectra featured by high spectral resolution may contain abundant spectral information, which may complicate the SOC estimation models [31,32]. Thus, it is necessary to establish new approaches to improve model parsimony.

Spectral variable selection aims to select the optimal variable subset from the full spectra [33–35]. This strategy can eliminate uninformative and noisy variables [31,32]. Several spectral variable selection methods have been proposed in previous studies. These methods can be categorized broadly into two categories [33]: (1) Methods that use regression coefficients, latent variables, and t-statistic of the full spectra, such as variable importance in projection (VIP) [34], uninformative variable elimination (UVE) [35], successive projections algorithm (SPA) [36], and competitive adaptive reweighted sampling (CARS) [37]; (2) methods that search for the optimal variable subset from the full spectra until the prediction error is minimum, such as genetic algorithm (GA) [38], simulated annealing (SA) [39], random frog (RF) [40], and ant colony optimization (ACO) [41]. In the first category, CARS is one of the most common methods to select soil spectral variables [33,42]. This algorithm is designed by the principle of ‘survival of the fittest’, which gradually eliminates unimportant variables when extracting the optimal variable subset according to the variable regression coefficient. In the second category, RF, proposed by Li et al., searches variable subset from global variables by reversible jump Markov Chain Monte Carlo technique [40]. This algorithm selects spectral variables by considering joint effects. It has been applied to soil total nitrogen spectra and plant spectra, showing good performance [43–45]. However, the effectiveness of this algorithm for SOC estimation by VIS-NIR spectroscopy is not yet clear. Therefore, the performance of CARS and RF is selected in this study.

In addition, spectral pretreatment is an important step in soil spectra analysis [46,47]. The raw full spectra may be influenced by instrument condition, measurement environment, and sample conditions. Pretreatment removes extraneous interference and enhances spectral information, which improves the performance of estimation models [48]. Common spectral pretreatments for VIS-NIR spectra include scattering correction, spectral derivatives, transformation, and range scaling [49]. Scatter correction,

including multiplicative scatter correction (MSC) and standard normal variate (SNV), could reduce the effects of soil particle size, surface roughness, and surface scattering. The first derivative (FD) and second derivative (SD) are typical methods of spectral derivatives, which could remove vertical offsets, linearly sloping baselines, and additive and multiplicative effects [50]. Log (1/R) is used to transform the reflectance spectra to absorption spectra, which can highlight the position with obvious absorption characteristics. Besides, it can increase the linearization between spectra and SOC to fit a better linear regression model. Mean centering (MC), belonging to range scaling methods, can eliminate the difference between each spectrum and improve the stability of the regression model [51]. It has been recognized that spectral pretreatment could influence the performance of VIS-NIR estimation models. However, the performance of different pretreatments with spectral variable selection algorithms remains unclear, and this necessitates further research.

This paper aimed to (1) explore the effect of CARS and RF spectral variable selection algorithms on the accuracy and parsimony of PLSR models to estimate SOC in anthropogenic soils; (2) compare spectral variable selection models with the full-spectrum model; and (3) test the performance of spectral pretreatments with spectral variable selection techniques.

## 2. Materials and Methods

### 2.1. Sampling Area and Soil Samples

The study area is located in the Jiangnan Plain, China. It is known as ‘Land of Fish and Rice’. The elevation changes from 22 m to 28 m. The study area is under a subtropical warm monsoon climate. The mean annual temperature is 16.10 °C, the mean annual precipitation is 1154 mm, and the mean annual relative humidity is 80%. The land-use types include cropland, woodland, and meadows. Cropland patches are highly fragmented, and some of them are close to settlements and various water bodies (breeding, ponds, irrigated canals, lakes, and rivers) [52]. Diverse land management practices are carried out in our study area according to our field survey. Intensive human activities have led to the heterogeneity of the relationship between VIS-NIR spectra and SOC [18].

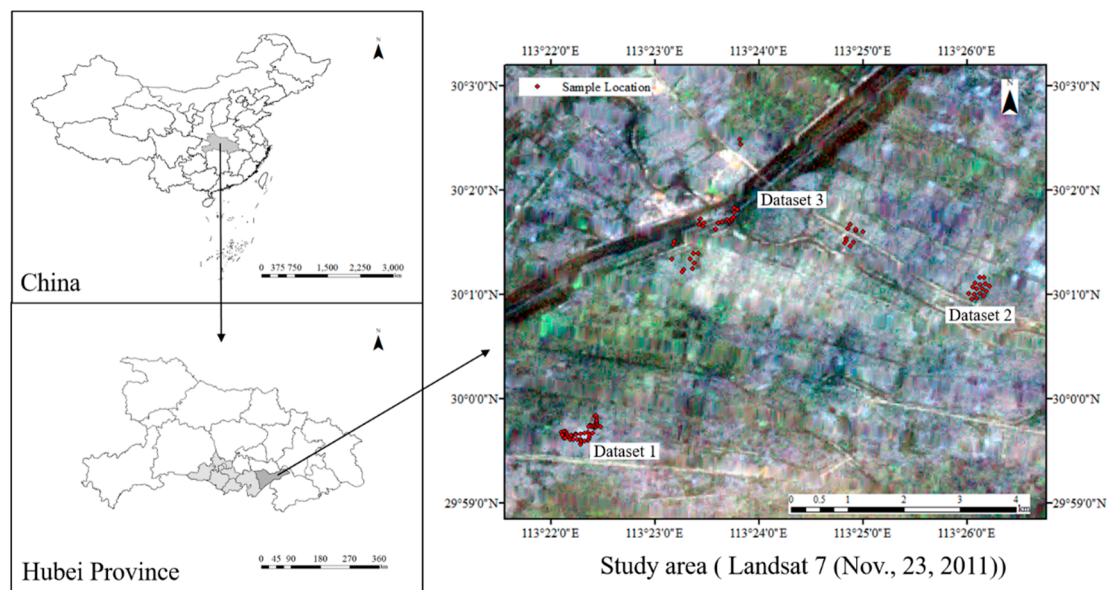
A total of 108 topsoil samples (0–15 cm) were collected from 20 December 2011 to 21 December 2011. Each soil sample was a mix of five soil subsamples, which were collected from the center and four corners in a square of 1 m<sup>2</sup> [53]. The geographical coordinates of these samples were recorded by a hand-held global positioning system, and the geographical distributions are shown in Figure 1. The total collected soil samples (Dataset 0) were divided into three datasets according to sampling locations, land use and land cover types (Dataset 1, Dataset 2, and Dataset 3, respectively). Samples of the three datasets were collected from three sites with different human activities on a small scale [18]. Samples of Dataset 1 was collected from cropland that was adjacent to a breeding pond. Dataset 2 was sampled from cropland that was surrounded by cropland. Dataset 3 included samples of various land-use types (cropland, artificial forest, meadows, and breeding ponds). These samples were put in sealed plastic bags with sampling sequence labels and then were sent to the laboratory at room temperature on 22 December 2011. After a principle components analysis and a 3 $\sigma$  standard, five outliers were discarded, and 103 samples were retained for further data analysis in this study.

### 2.2. VIS-NIR Spectral Measurement and SOC Analysis

In the laboratory, soil samples were air-dried at 20–30 °C for 1 week, then ground, and passed through a 2-mm sieve [54]. An ASD FiledSpec 3 portable spectro-radiometer with a spectral range of 350–2500 nm, and a spectral resolution of 1 nm was used to scan soil samples in a dark room to avoid stray light interference. All samples were put separately in dishes with a 20-cm diameter. A halogen lamp placed at 30 cm distance and an angle of 45° was used to illuminate soil samples. The detection fiber probe was placed vertically to soil samples at 12 cm distance. A white Spectralon panel was used to calibrate the spectrometer before measuring spectrum of the first soil sample and repeated every six soil samples [53]. A total of 10 scans were recorded for each soil sample, which were averaged in

one sample spectrum [55]. Through these procedures, the reflectance spectra of the 108 samples were obtained. The spectra in the range of 350–399 nm and 2450–2500 nm were removed due to serious noises. The remained spectra (400–2449 nm) were further resampled to 10 nm to extract 205 wavebands.

The SOC content was measured by wet oxidation at 180 °C with a mixture of potassium dichromate and sulfuric acid [18]. It should be noted that the oxidation of active organic carbon by this approach is incomplete, which underestimates the SOC content. A “standardized” corrective factor ranging from 1.10 to 1.40 could be used in practice [56]. In our study, we used the “raw” SOC content without using a “standardized” corrective factor. This allowed comparing the results of our study with other studies that also use the “raw” SOC content.



**Figure 1.** Location of the study area and soil sampling.

### 2.3. Spectral Pretreatment

We selected five commonly used pretreatments for raw spectra, which included FD, Log (1/R), MC, MSC, and SNV. All the spectral pretreatments were performed in Matlab (R2014b, MathWorks, Inc., Natick, MA, USA) with the PLS toolbox (Version 7.9.3, Eigenvector Research, Inc., Wenatchee, WA, USA).

### 2.4. Spectral Variable Selection

CARS and RF were chosen to achieve spectral variable selection in our study. CARS is designed on the principle named ‘survival of the fittest’ [16,37]. When using this algorithm to select the spectral variables, each spectral waveband is treated as an individual. The adaptive individuals are left behind as the final spectral variable subset, while the maladaptive individuals are eliminated. The scheme of the CARS algorithm can be summarized as five steps: (1) utilize the Monte Carlo algorithm to randomly choose  $n$  samples and build Partial least squares regression (PLSR) models for these subsets; (2) retain variables with high regression coefficients and remove variables with low regression coefficients through an exponentially decreasing function and adaptive reweighted sampling algorithm; (3) build PLSR model and compute the root mean square error of cross validation ( $RMSE_{cv}$ ) with the retained variables as a new variable subset; (4) repeat steps 1–3 for  $N$  runs to obtain  $N$  new variable subsets; and (5) choose the new subsets with the lowest  $RMSE_{cv}$  as the optimal variable subset [41]. The run times ( $N$ ) was set to 50 in this study.

Random frog (RF) is also a variable selection algorithm, which is inspired by the reversible jump Markov Chain Monte Carlo (RJMCMC) technique [40]. It is mathematically easier and more computationally efficient than RJMCMC. RF searches for the best variable subset from a full variable



set by considering the interaction between spectral variables. Five steps are needed to realize this algorithm: (1) randomly initialize variable subset  $K0$  consisting of  $Q$  variables; (2) obtain new variable subset  $K^*$  using the way of the normal distribution, then update the variable subset  $K0$  as  $K1$  through the cross-validated misclassification errors of  $K0$  and  $K^*$ ; this step needs to be iterated many times; (3) compute each variable selection probability after the iterations finish, and variables with same selection probability are selected as variable subsets; (4) build PLSR model with each variable subset, and compute  $RMSE_{cv}$ ; and (5) choose the variable subset with the lowest  $RMSE_{cv}$  as the optimal variable subset.  $X$  ( $m \times n$ ),  $Y$  ( $n \times 1$ ),  $N$ , and  $Q$  are the key input parameters for this algorithm, where  $X$  is the spectral variables, and  $Y$  is the SOC content, in which  $m$  is the number of samples, and  $n$  is the number of spectral variables.  $N$  is the number of iterations, and  $Q$  is the number of variables consisting of the initialized variable subset [44]. The number of iterations ( $N$ ) was set to 50 in this study. In this study, CARS and RF were performed in Matlab (R2018b, MathWorks, Inc., Natick, MA, USA) with the libPLS toolbox (Version 1.98), which was available at <http://www.libpls.net/download.php>.

## 2.5. Model Calibration and Validation

The 103 samples were divided into calibration and validation datasets using the concentration gradient method [16]. According to the SOC values, the 103 samples were sorted in ascending order. Then, the 103 samples were divided into a total of 35 groups. The last group included one sample, and the other groups had three samples. In each group, the first and third samples were selected and added into the calibration dataset, whereas the second sample was added into the validation dataset. Thus, the calibration dataset contained 69 samples, and the validation dataset contained 34 samples.

Partial least squares regression (PLSR) was chosen to develop regression models between SOC and VIS-NIR spectra variables, which is proposed by Wold et al. [57]. It creates a linear regression model by projecting the independent and dependent variables into a new space [58]. This method integrates the advantages of principal component analysis, canonical correlation analysis and linear regression. It can overcome the multicollinearity problem and improve model performance. In this study, the PLSR models were built on the calibration dataset and the performance of the PLSR models was evaluated on the validation dataset. The coefficient of determination ( $R^2$ ), the root mean squared error ( $RMSE$ ) and the residual prediction deviation ( $RPD$ ) were calculated using Equations (1)–(3) [57]. A desirable PLSR estimation model should have high  $R^2$  and  $RPD$  with a low  $RMSE$  on the validation dataset.

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

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

$$RPD = \frac{SD}{RMSE_P} \quad (3)$$

where  $i$  is the  $i$ th soil sample,  $n$  is the number of soil samples,  $\bar{y}$  is the mean of the measured SOC content,  $y_i$  is the measured SOC content for the  $i$ th soil sample,  $\hat{y}_i$  is the predicted SOC content for the  $i$ th soil sample,  $SD$  is the standard deviation of the measured SOC content of the validation dataset, and  $RMSE_P$  is the root mean square error of the validation dataset.

## 3. Results

### 3.1. Statistical Description of Soil Samples

Table 1 describes the statistics of soil samples for the total, calibration and validation datasets. The SOC content of the total and the calibration datasets ranged from 2.35 g/kg to 33.95 g/kg, whereas the range of SOC content of the validation dataset was smaller. This indicated that the calibration dataset covered representative samples for successful model calibration establishment. The coefficient of

variation (CV) values of the total, calibration and validation datasets were 40%, 40%, and 39%, respectively, indicating moderate variability. The coefficient of skewness (CS) values were  $-0.04$ ,  $0.04$ , and  $-0.23$ , respectively, whereas the coefficient of kurtosis (CK) values were  $2.32$ ,  $2.46$ , and  $1.93$ , respectively. It can be reported that these datasets were approximately of normal distribution with small dispersion.

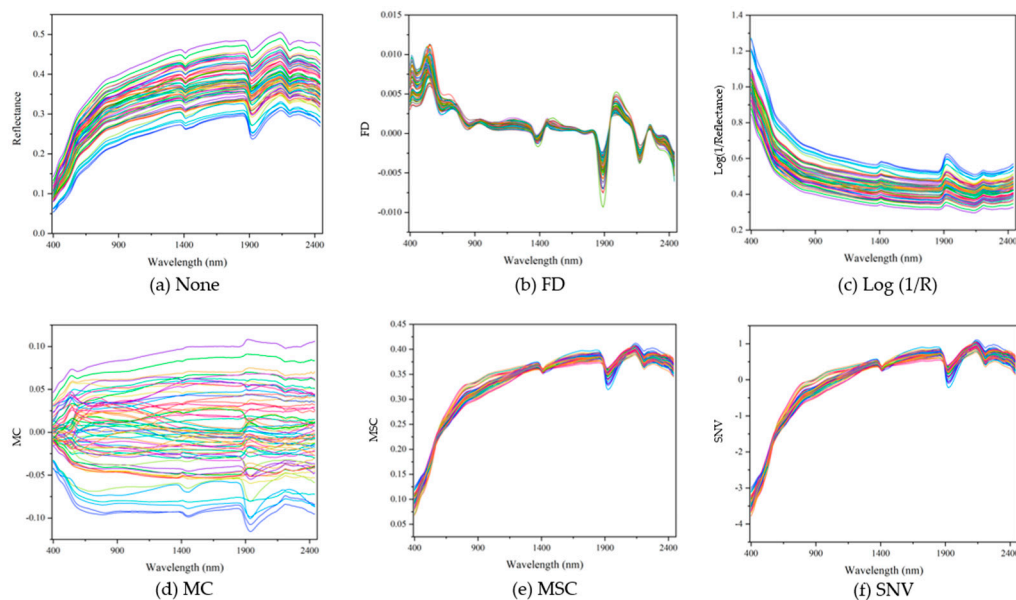
**Table 1.** Descriptive statistics of soil organic carbon (SOC) of soil samples (Dataset 0).

Samples	N <sup>a</sup>	SOC (g/kg)			SD <sup>d</sup>	CV <sup>e</sup>	CS <sup>f</sup>	CK <sup>g</sup>
		Min <sup>b</sup>	Max <sup>c</sup>	Mean				
Total	103	2.35	33.95	16.05	6.35	40%	$-0.04$	2.32
Calibration	69	2.35	33.95	16.14	6.46	40%	$0.04$	2.46
Validation	34	3.30	26.23	15.85	6.20	39%	$-0.23$	1.93

<sup>a</sup> Sample numbers; <sup>b</sup> Minimum; <sup>c</sup> Maximum; <sup>d</sup> Standard deviation; <sup>e</sup> Coefficient of variation; <sup>f</sup> Coefficient of skewness; <sup>g</sup> Coefficient of kurtosis.

### 3.2. Raw Spectra and Pretreated Spectra

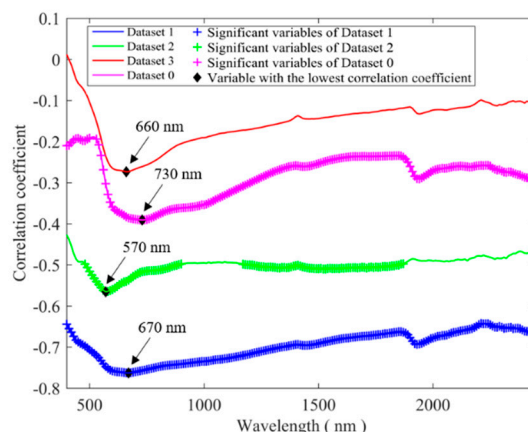
Figure 2 shows raw spectra and pretreated spectra with different pretreatments. In Figure 2a, raw spectra had a similar shape for all samples, whose reflectance rose rapidly before 850 nm and slowly after 850 nm. There were three absorption peaks with different depths distinguishable in the near-infrared region, which are attributed to the hydroxyl group of free water (1410 nm and 1920 nm) and the Al-OH group of clay minerals (2210 nm) [59]. After FD, the most spectral values tended to approach zero, highlighting the bands where the curvature of raw spectra changed greatly. Spectra after Log (1/R) pretreatment showed typical decline in absorption at different rates with the range of 400–2449 nm. After MC, spectral values were between  $-0.15$  and  $0.15$ . There was still a significant fluctuation at around 1410 nm, 1920 nm, and 2210 nm. The spectra after MSC and SNV had similar shapes, but with different ranges.



**Figure 2.** The raw and pretreated soil spectra. (a) None: raw spectra; (b) FD: the spectra after first derivative; (c) Log (1/R): the absorption spectra; (d) MC: the spectra after mean centering; (e) MSC: the spectra after multiplicative scatter correction; and (f) SNV: the spectra after standard normal variate.

### 3.3. Correlation Analysis

The differences in the correlation coefficient curves reveal heterogeneous relationships between raw VIS-NIR spectra and SOC (Figure 3). In Figure 3, blue '+', green '+', and magenta '+' symbols refer to locations of VIS-NIR spectral variables having significant correlations for Dataset 1, Dataset 2, and Dataset 0, respectively (at a significance level of 0.05). It was revealed that SOC had a significantly negative correlation with raw VIS-NIR spectra in the region of 400–2449 nm for Dataset 1 and Dataset 0. The spectral variables with significant negative correlations distributed in the region of 480–900 nm and 1170–1870 nm for Dataset 2, whereas no significant correlations were observed for Dataset 3.

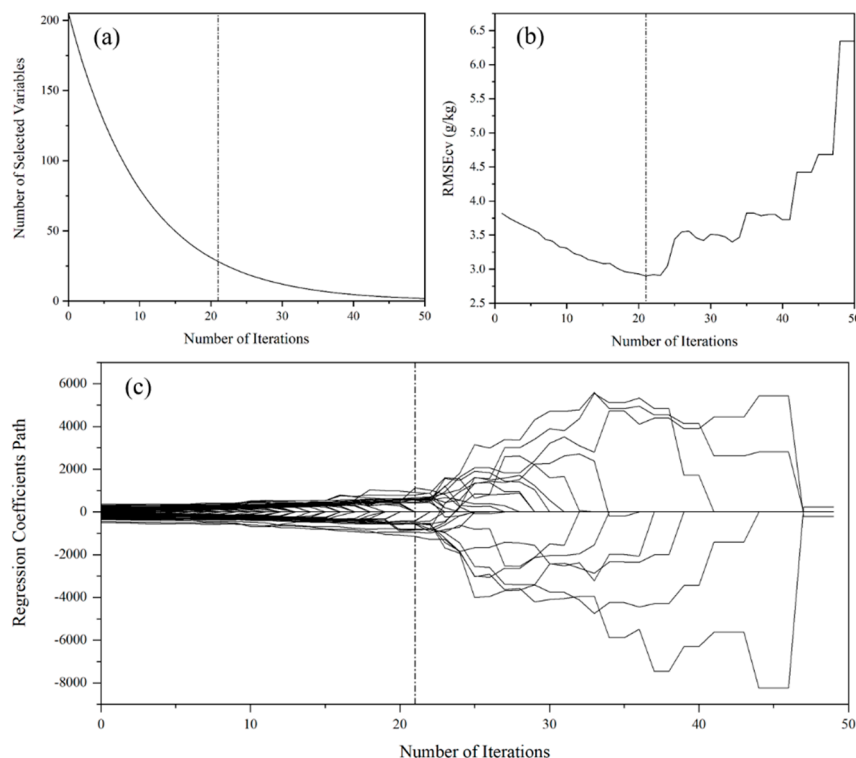


**Figure 3.** Correlation coefficient curves calculated between the raw visible and near-infrared (VIS-NIR) spectra and soil organic carbon (SOC) for four datasets. The blue line, green line, red line, and magenta line refer to correlation coefficient curves for Dataset 1, Dataset 2, Dataset 3, and Dataset 0, respectively. The blue '+', green '+', and magenta '+' symbols refer to locations of VIS-NIR spectral variables having significant correlation for Dataset 1, Dataset 2, and Dataset 0, respectively (at a significance level of 0.05). The '◆' symbol refers to the location of spectral variables having the lowest correlation coefficient.

Dataset 1 had the strongest correlations among these four datasets. The absolute correlation coefficients slowly decreased as the wavelength increases after 670 nm. The correlation coefficients of Dataset 2 had a faint change in the spectral range of 870–2449 nm. Dataset 0 was the combination of Dataset 1, Dataset 2 and Dataset 3. Absolute correlation coefficients for Dataset 0 increased in the range of 400–570 nm, which was a different trend compared to Dataset 1, Dataset 2 and Dataset 3. The highest absolute correlation coefficients of Dataset 1 (670 nm), Dataset 2 (570 nm), Dataset 3 (660 nm), and Dataset 0 (730 nm) were of different magnitudes. This provides vivid evidence that a heterogeneous relationship exists between VIS-NIR spectra and SOC for soils with intensive human activities.

### 3.4. Spectral Variable Selection

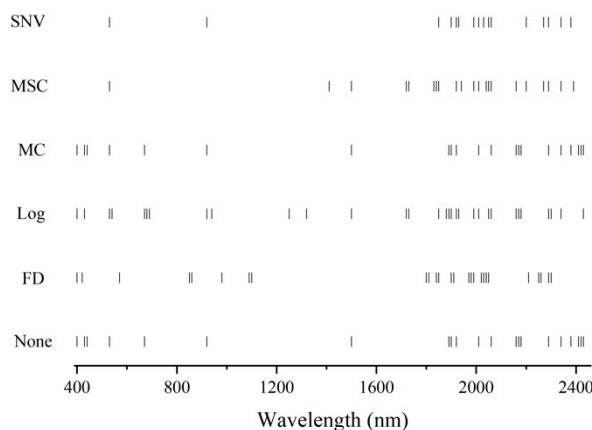
Figure 4 shows details of spectral variables selected by CARS, taking the Log (1/R) spectra as an example. Figure 4a shows the number of selected spectral variables during 50 iterations. As the number of iterations increases, the selected spectral variables first decreased sharply and then gradually. Figure 4b shows the change of 5-fold RMSEcv during the 50 iterations. As the number of iterations increased, the RMSEcv value first declined and then ascended. At 21 iterations, RMSEcv reached a minimum. Figure 4c shows the regression coefficient path of all spectral variables (regression coefficient path: the changing trend of regression coefficient values during the 50 iterations). At iteration one, the regression coefficients of spectral variables were similar. With the increase of iterations, the regression coefficient of some spectral variables gradually increased, while they remained zero for others. According to Figure 4a–c, at iteration 21, the selected spectral variables were the optimal subset that helped to achieve the best model performance.



**Figure 4.** Competitive adaptive reweighted sampling (CARS) variable selection of Log (1/R) spectra: (a) the number of sampled variables; (b) 5-Fold root mean squared error of cross-validation ( $RMSE_{cv}$ ) values; and (c) regression coefficient path of each spectral variable during the 50 iterations.

For the other spectral pretreatments, the number of selected spectral variables also decreased with iterations. Different iterations were required to reach the minimum  $RMSE_{cv}$  for spectra with different pretreatments (e.g., 25, 23, 25, 25, and 28 for the pretreatments of None, FD, MC, MSC, and SNV, respectively).

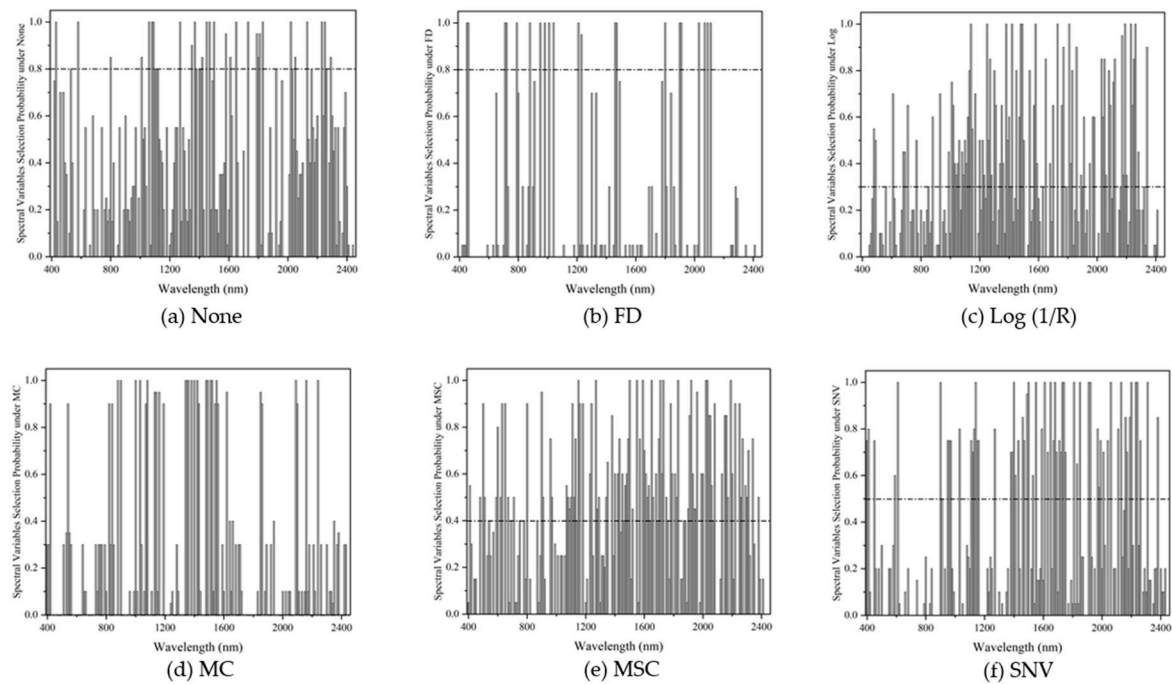
Figure 5 shows the spectral variables selected by the CARS method with different spectral pretreatments. A total of 21, 26, 31, 21, 21, and 16 spectral variables were selected as the optimal spectral subset for the pretreatments of None, FD, Log (1/R), MC, MSC, and SNV, respectively. Most of the selected spectra variables were concentrated in the 1800–2449 nm spectral range. Only one spectral variable was selected in the range of 400–780 nm for MSC and SNV. No spectral variable was selected in the range of 1200–1800 nm for FD and SNV, and neither in the range of 780–1200 nm for MSC.



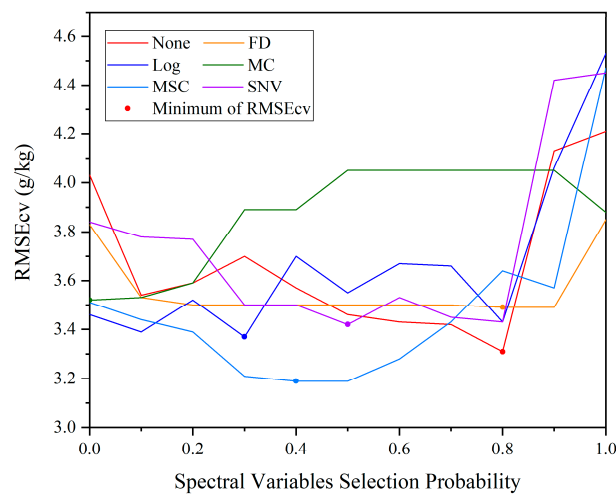
**Figure 5.** The distributions of spectral variables selected by competitive adaptive reweighted sampling (CARS) with different spectral pretreatments.



Figure 6 shows the probability that each spectral variable is repeatedly selected during the 50 iterations. With the increase of selection probability, the number of selected spectral variables reduced. For different spectral pretreatments, each spectral variable had different selection probability. The number of selected spectral variables was the least when spectral pretreatment was FD. Figure 7 shows the change of 5-Fold RMSE<sub>cv</sub> in different spectral variable selection probability. Spectral variables were selected as the optimal subset with the lowest 5-Fold RMSE<sub>cv</sub>. According to Figure 7, RMSE<sub>cv</sub> had minimum values when the probability was 0.8, 0.8, 0.3, 0, 0.4, and 0.5 for the spectral pretreatments of None, FD, Log (1/R), MC, MSC, and SNV, respectively.

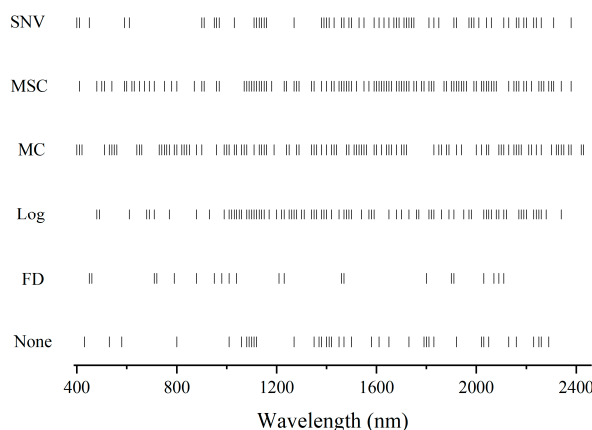


**Figure 6.** Spectral variables selection probability by Random Frog (RF) with different spectral pretreatments. (a) probability without pretreatment; (b) probability with first derivative (FD); (c) probability with Log (1/R); (d) probability with mean centering (MC); (e) probability with multiplicative scatter correction (MSC); and (f) probability with standard normal variate (SNV).



**Figure 7.** Five-fold root mean square error of cross-validation (RMSE<sub>cv</sub>) for different spectral variable selection probability by Random Frog (RF) shown for different spectral pretreatments.

Figure 8 shows spectral variables selected by RF with different spectral pretreatments. A total of 39, 21, 83, 101, 106, and 63 spectral variables were selected as the optimal subset for the pretreatments of None, FD, Log (1/R), MC, MSC, and SNV, respectively. More spectral variables were selected in the 400–780 nm spectral range for MC and MSC spectra pretreatments.



**Figure 8.** The distributions of spectral variables selected by Random Frog (RF) with different spectral pretreatments.

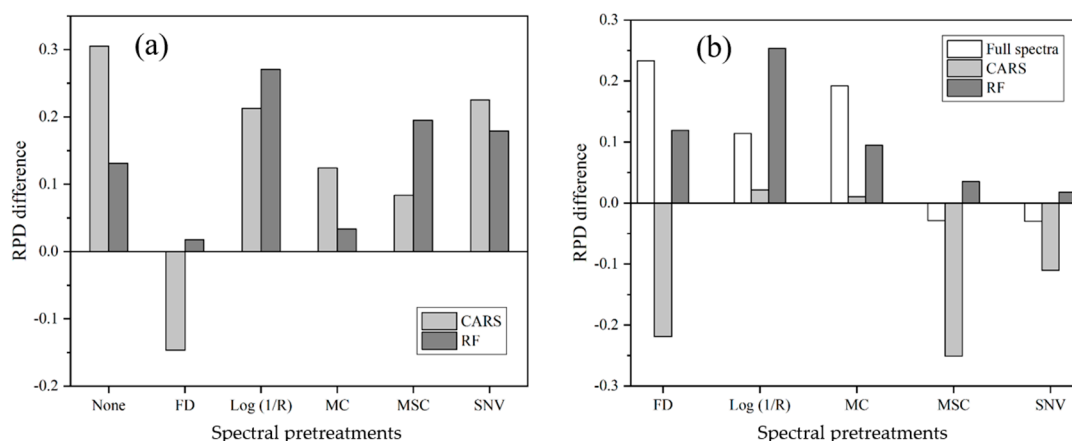
### 3.5. Accuracy of Estimation after Different Pretreatment and Variable Selection Techniques

To evaluate the effectiveness of different variable selection methods and the influence of different pretreatments tested, PLSR was used to establish a series of SOC models, whose results are shown in Table 2 and Figure 9.

**Table 2.** Accuracies of soil organic carbon (SOC) estimation using full-spectrum-based partial least squares regression (PLSR) models, competitive adaptive reweighted sampling (CARS)-based PLSR models, and random frog (RF)-based PLSR models after different spectral pretreatments.

Spectral Variable Selection	Spectral Pretreatments	N <sup>a</sup>	LVs <sup>b</sup>	Calibration Dataset		Validation Dataset		RPD	$\overline{RPD}$ <sup>c</sup>
				R <sub>c</sub> <sup>2</sup>	RMSE <sub>c</sub>	R <sub>p</sub> <sup>2</sup>	RMSE <sub>p</sub>		
Full Spectra	None	205	9	0.79	2.93	0.70	3.60	1.72	1.81
	FD	205	7	0.78	3.01	0.80	3.17	1.96	
	Log(1/R)	205	11	0.86	2.44	0.76	3.37	1.84	
	MC	205	10	0.86	2.36	0.75	3.24	1.92	
	MSC	205	8	0.78	3.02	0.70	3.66	1.70	
	SNV	205	8	0.78	3.02	0.70	3.66	1.69	
CARS	None	21	8	0.85	2.45	0.78	3.05	2.03	1.94
	FD	26	7	0.85	2.44	0.73	3.42	1.81	
	Log(1/R)	31	8	0.84	2.53	0.81	3.02	2.05	
	MC	21	8	0.87	2.35	0.78	3.04	2.04	
	MSC	21	6	0.79	2.91	0.77	3.49	1.78	
	SNV	16	6	0.83	2.66	0.77	3.23	1.92	
RF	None	39	10	0.83	2.61	0.72	3.34	1.86	1.94
	FD	21	14	0.84	2.53	0.76	3.14	1.97	
	Log(1/R)	83	11	0.86	2.42	0.83	2.94	2.11	
	MC	101	10	0.85	2.45	0.77	3.18	1.95	
	MSC	106	11	0.89	2.16	0.76	3.28	1.89	
	SNV	63	8	0.81	2.83	0.77	3.31	1.87	

<sup>a</sup> Number of selected spectral variables; <sup>b</sup> Number of latent variables; <sup>c</sup> Mean of ratio of prediction deviation (RPD).



**Figure 9.** (a) Ratio of prediction deviation (RPD) difference between competitive adaptive reweighted sampling (CARS)/random frog (RF) and full spectrum SOC models after the same spectral pretreatments (spectral pretreatments include non-pretreatment (None), first derivative (FD), Log (1/R), mean centering (MC), multiplicative scatter correction (MSC), and standard normal variate (SNV)); (b) RPD difference between non-pretreated and pretreated SOC models after the same variable selection algorithms (variable selection algorithms include full spectrum, CARS, and RF).

Both CARS and RF could enhance the performance of PLSR models (except for the combination of CARS and FD, Figure 9a), compared to the full-spectrum models. The  $\overline{RPD}$  of the full-spectrum models was 1.81, while the  $\overline{RPD}$ s of CARS and RF models were of equal value of 1.94 (Table 2). The best  $R_p^2$  in each spectral variable selection category slightly increased from 0.80 (Full spectrum) to 0.81 (CARS) and 0.83 (RF), and RPD increased from 1.96 (Full spectra) to 2.05 (CARS) and 2.11 (RF).

Spectral pretreatments have different effects on the full-spectrum PLSR models and the PLSR models obtained after spectral variable selection (Table 2 and Figure 9b). For the full-spectrum PLSR models, FD, Log (1/R) and MC had positive effects, as  $R_p^2$  increased from 0.70 to 0.80;  $RMSEP$  decreased from 3.60 g/kg to 3.17 g/kg; and RPD increased from 1.72 to 1.96. MSC and SNV had negative effects, as  $R_p^2$  remained 0.70;  $RMSEP$  increased from 3.60 g/kg to 3.66 g/kg; and RPD decreased from 1.72 to 1.69. Full-spectrum PLSR model with FD had the largest RPD difference (0.23), and that with MSC and SNV had the lowest RPD difference (−0.03).

For the CARS + PLSR models, the spectral pretreatments of Log (1/R) and MC enhanced the PLSR model performance, but the spectral pretreatments of FD, MSC and SNV weakened PLSR model performance. The best  $R_p^2$ ,  $RMSEP$  and RPD values were 0.81, 3.02 g/kg and 2.05, respectively. The worst  $R_p^2$ ,  $RMSEP$  and RPD values were 0.73, 3.42 g/kg and 1.81, respectively. CARS + PLSR model with MSC had the largest RPD difference (−0.25), and that with MC had the lowest RPD difference (0.01).

For the RF + PLSR models, all spectral pretreatments had positive effects.  $R_p^2$  increased from 0.72 to 0.83;  $RMSEP$  decreased from 3.34 g/kg to 2.94 g/kg; and RPD increased from 1.86 to 2.11. The RF + PLSR model with Log (1/R) had the largest RPD difference (0.23), and that with SNV had the lowest RPD difference (−0.03).

Therefore, the Log (1/R) was the best pretreatment for spectral variable selection. Log (1/R) + CARS/RF + PLSR models provided the best performance (Log (1/R) + CARS + PLSR model:  $R_p^2 = 0.81$ ,  $RMSEP = 3.02$  g/kg, and  $RPD = 2.05$ , Log (1/R) + RF + PLSR model:  $R_p^2 = 0.83$ ,  $RMSEP = 2.94$  g/kg, and  $RPD = 2.11$ ).

The number of selected spectral variables after RF was larger than that after CARS. For RF, the minimum and maximum numbers of selected spectral variables were 21 and 106, respectively, whereas there were 16 and 31 for CARS. Besides, the number of latent variables after RF was larger than that after CARS.

## 4. Discussion

### 4.1. The Effect of Spectral Variable Selection Techniques on Model Accuracy

Numerous studies have proven that SOC estimation by VIS-NIR spectra is efficient and accurate. Nevertheless, it remains a challenge to apply VIS-NIR spectroscopy to estimate SOC in anthropogenic soils that are characterized by high heterogeneity in the relationship between VIS-NIR spectra and SOC. The study area, Jiangnan plain, has been under a long-term period of human activities with a highly fragmented landscape [52,60]. Our study reported that there was strong heterogeneity in the relationship between SOC and VIS-NIR spectra. With soil samples collected from the study area, we investigated the effect of two spectral variable selection techniques to improve the performance of the PLSR models in SOC estimation in anthropogenic soils. Our results demonstrated that spectral variable selection could improve the accuracy of PLSR models, wherein  $\overline{RPD}$  increased by 0.13 (without spectral variable selection: the best  $R^2 = 0.80$ ,  $RPD = 1.96$  and  $\overline{RPD} = 1.81$ ; with spectral variable selection: the best  $R^2 = 0.83$ ,  $RPD = 2.11$  and  $\overline{RPD} = 1.94$ ). This is because that spectral variable selection could eliminate unimportant information, reserve relevant information, and reduce spectral collinearity [37]. The performance of the proposed spectral variable selection was comparable to other strategies that aimed to improve the accuracy of PLSR models for anthropogenic soil. Liu et al. compared a variety of sample selection algorithms, which aimed to develop a representative calibration dataset for SOM estimation [18]. The best  $RPD$  achieved in their study was lower than the  $\overline{RPD}$  in this study. Liu et al. further combined the Kennard–Stone algorithm and spectral pretreatment to choose representative calibration samples, and achieved an  $\overline{RPD}$  of 1.85, which was still poorer than that obtained in the current study [59]. Wang et al. proposed the MVARC-R-KS method to select representative calibration samples (not spectral variables as in the current study), which has resulted in good accuracy of PLSR models [61]. They reported that the best  $RPD$  was 1.81, which was also lower than the  $\overline{RPD}$  in this study. These mentioned strategies mainly focus on the selection of representative calibration samples to improve the accuracy of PLSR models, while our strategies focus on the selection of representative spectral variables. A combination of these two strategies to further improve the performance of PLSR models in SOC estimation could be explored in future research.

We also tested the performance of spectral pretreatments with spectral variable selection techniques. The PLSR models performed differently using different selected spectral variables after different spectral pretreatments. For the CARS-based models, Log (1/R) and MC enhanced the accuracy of PLSR models, while FD, MSC and SNV led to accuracy deterioration. FD has not improved the estimation performance of the PLSR model. This might be due to the lack of variables in the spectral region of 1200–1800 nm, a spectral range that contains relevant overtones and combinations of fundamental bonds of O–H and C–O groups that are associated with SOC. The amplification of noise may also hinder useful signals [62]. MSC has failed to improve model performance as only one variable was selected in the visible region of 400–780 nm, which is an important region associated with soil color variation in the blue band around 450 nm. The darker the soil color, the higher the SOC content [63]. This single selected variable may not have accounted for sufficient information to capture variation in soil darkness that could be linked with variation in SOC. The number of selected spectral variables with SNV was large than that with MSC, which explains why the PLSR model with SNV outperforms that with MSC. For the RF-based models, all pretreatments enhanced the accuracy of PLSR models. It could be attributed to the fact that RF could select important spectral variables as comprehensively as possible while reducing collinearity between spectral variables. Among them, Log (1/R) + CARS/RF + PLSR models provided the best estimation accuracy. Since SOC absorbs energy at specific frequencies in VIS-NIR, Log (1/R) enhanced its absorption feature [64].

It should be noted that the effects of spectral pretreatments are determined by the quality of raw spectra. In other words, spectral pretreatments would be necessary in the case that raw spectra are influenced by variable soil physical conditions and by the surrounding environment during measurement (e.g., temperature, ambient light). In this study, however, all the samples were air-dried,

ground and passed through a 2-mm sieve, and VIS-NIR spectra were measured in a well-controlled environment. Therefore, both CARS and RF could develop satisfactory SOC estimation models without spectral pretreatments. Meanwhile, this study provided that Log (1/R) should be adopted to further improve the accuracy of SOC estimation models.

#### 4.2. The Effect of Spectral Variable Selection Techniques on Model Parsimony

Model parsimony is an important issue that has been less investigated in previous studies. With the demand for advanced regression algorithms, fewer spectral variables are desired for SOC estimation and mapping when using spectra data. This study demonstrated fewer spectral variables were needed to build PLSR models by CARS and RF algorithms, as the numbers of selected variables were 31 and 83 for CARS and RF, respectively.

The search mechanism of spectral variables by CARS and RF algorithms is different, which results in a different number of selected spectral variables and their distributions. CARS selects spectral variables that have the largest regression coefficient through continuous iterations. In our study, the selected spectral variables were mainly concentrated in the spectral region around 540 nm and 1800–2400 nm. This result was similar to that of Vohland et al. and Hong et al. [16,33,42]. Vohland et al. selected variables at 1875 nm, 2205–2220 nm, and 2330–2345 nm to successfully estimate SOC content in the Eifel-Moselle-Hunsrück region. Hong et al. reported spectral variables at 400–800 nm and 2000–2400 nm to be associated with SOC. It should be noted that sensitive spectral variables for SOC may differ from one dataset to another. This explains why the selected spectral variables by CARS are not the same in different studies. The number of the selected spectral variables by RF is larger than that by CARS, and the distribution of its selected spectral variables is sparser. It can be explained by the fact that RF globally searches the optimal variable combination based on random frog leaping. The selected spectral variables by RF were mainly distributed around 430 nm, 610 nm, 800 nm, 1000 nm, 1100 nm, 1200 nm, 1420 nm, 1500 nm, 1800 nm, 1920 nm, 2000 nm, 2100 nm, 2200 nm, and 2350 nm.

In summary, the selected spectral variables in the visible region by both methods are mainly concentrated in the locations of 400 nm, 530 nm and 610 nm. These wavelengths are associated with variation in chromophores and the darkness of humic acid (e.g., due to blue absorption band at 450 nm and perhaps red color absorption band at 680 nm) [65]. Spectral variables in the near-infrared region are chosen because some fundamental vibrational bonds are associated with SOC. These bonds mainly include C–H, N–H, C–O, C=O, O–H, and Al–OH [54,66]. Table 3 shows the possible assignments of fundamental bonds, absorption wavelength, and related soil constituents for the main selected spectral variables in the near-infrared region by RF and CARS. Some selected spectral wavelengths in this study do not coincide with possible wavelengths of fundamental vibrational bonds. It could be attributed to slight shifts in wavelength locations due to inharmonic molecules vibrations [54].

**Table 3.** Possible assignments of fundamental bonds, absorption wavelength, and related soil constituent for the selected spectral variables in the near-infrared region by competitive adaptive reweighted sampling (CARS) and random frog (RF) [54,66].

Locations of Selected Spectral Variables (nm)	Possible Fundamental Bonds	Possible Wavelength (nm)	Possible Related Soil Constituents
800	C–H	825	Organics (aromatics)
1000	N–H	1000	Organics (amine)
1100	C–H	1100	Organics (aromatics)
1200	C–H	1170	Organics (Alkyl)
1420	O–H		asymmetric-symmetric doublet)
1500	C–O	1380	Water
1800	C–H	1524	Organics (amides)
1920	O–H	1754	Organics (Alkyl)
2000	C–O		asymmetric-symmetric doublet)
2100	N–H	1915	Water
2200	Al–OH	2033	Organics (amides)
2350	C–O	2060	Organics (amine)
		2230	Clay minerals
		2381	Organics (Carbohydrates)



### 4.3. The Implication of the Proposed Strategy

Accuracy and parsimony of PLSR models could be improved by the proposed modeling approach investigated in this study, but could not be optimal at the same time. The more spectral variables, the more complex PLSR models and vice versa. However, the inclusion of a larger number of input spectral variables could be redundant, while a smaller number of spectral variables could cause loss of significant spectral information related to SOC. Both cases will result in low accuracy of PLSR models. An optimal number of selected spectral variables could balance the model parsimony and accuracy. Nevertheless, an optimal number of spectral variables is difficult to determine and varies from one dataset to another. Therefore, it is essential to search for the best pretreatment to obtain the most appropriate number of spectral variables. For the CARS algorithm, the PLSR model achieved the best accuracy when a total of 31 spectral variables were selected after the Log (1/R) spectra pretreatment. The number of selected spectral variables were smaller with other spectral pretreatments, which resulted in deteriorated accuracy of PLSR models due to the exclusion of significant variables related with SOC. For the RF algorithm, selected spectral variables with Log (1/R), MC, MSC, and SNV were of similar distribution. The PLSR model with Log (1/R) outperformed the models using the other three spectral pretreatments. This is possibly because of the redundant information that resulted from the larger number of selected spectral variables obtained with MC and MSC compared to that with Log (1/R). The number of spectral variables selected with SNV was smaller than that with Log (1/R), which may indicate the loss of important spectral information by SNV. Since the Log (1/R) has resulted in selecting an appropriate number of spectral variables, the best PLSR models could be established for SOC estimation using both spectral variable selection techniques (RF:  $R^2 = 0.83$  and  $RPD = 2.05$ ; CARS:  $R^2 = 0.81$ , and  $RPD = 2.11$ ).

## 5. Conclusions

In this study, competitive adaptive reweighted sampling (CARS) and random frog (RF) algorithms in combination with five different spectral pretreatments were used to select spectral variables, which were used as input in partial least squares regression (PLSR) to estimate soil organic carbon (SOC) in the Jiangnan Plain of China. According to our results, the following conclusions can be made: (i) Both spectral variable selection algorithms (e.g., CARS and RF) could improve the accuracy and parsimony of PLSR models; (ii) the accuracy of the PLSR model obtained after RF is better than that with CARS, although the parsimony of the latter model was worse (the best models with RF:  $R^2 = 0.83$ ,  $RPD = 2.11$ , and the number of spectral variables = 83; the best models with CARS:  $R^2 = 0.81$ ,  $RPD = 2.05$ , and the number of spectral variables = 31); (iii) the effects of spectral pretreatments vary among spectral variable selection algorithms. All FD, Log (1/R), MC, MSC, and SNV could improve the accuracy of PLSR models with RF, whereas only Log (1/R) and MC could slightly improve the accuracy of PLSR models with CARS; and (iv) appropriate number and distribution of spectral variables could be selected by Log (1/R) after both CARS and RF.

Although our study improved the accuracy and parsimony of PLSR models for SOC estimation using two spectral selection algorithms in anthropogenic soil, laboratory non-imaging sensors were adopted. Future researches should focus on airborne and satellite sensors to explore their potential for estimation and mapping topsoil SOC.

**Author Contributions:** All of the authors contributed to the study. Conceptualization, L.X., Y.H., and Y.C.; methodology, L.X., and Y.H.; software, L.X.; validation, L.X.; formal analysis, L.X., and Y.W.; investigation, L.X., and Y.W.; resources, Y.C.; data curation, Y.L. (Yi Liu), L.G., Q.J., T.S., T.F., Y.L. (Yaolin Liu), and Y.C.; writing—original draft preparation, L.X.; writing—review and editing, L.X., Y.H., A.M.M., and Y.C.; visualization, L.X., and Y.W.; supervision, Y.C.; project administration, Y.C.; funding acquisition, Y.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Natural Science Foundation of China (grant numbers: 41771440) and the Fundamental Research Funds for the Central Universities (grant numbers: 2042020kf0201).

**Acknowledgments:** The authors thank the editors and the reviewers for their constructive comments.

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

## References

- Schmidt, M.W.; Torn, M.S.; Abiven, S.; Dittmar, T.; Guggenberger, G.; Janssens, I.A.; Kleber, M.; Kögel-Knabner, I.; Lehmann, J.; Manning, D.A.C.; et al. Persistence of Soil Organic Matter as an Ecosystem Property. *Nat. Cell Biol.* **2011**, *478*, 49–56. [\[CrossRef\]](#)
- Wiesmeier, M.; Urbanski, L.; Hobbey, E.; Lang, B.; Von Lützow, M.; Marin-Spiotta, E.; Van Wesemael, B.; Rabot, E.; Ließ, M.; Garcia-Franco, N.; et al. Soil Organic Carbon Storage as a Key Function of Soils—A Review of Drivers and Indicators at Various Scales. *Geoderma* **2019**, *333*, 149–162. [\[CrossRef\]](#)
- Meuser, H. Anthropogenic Soils. In *Contaminated Urban Soils*; Meuser, H., Ed.; Springer: Dordrecht, The Netherlands, 2010; pp. 121–193.
- Dazzi, C.; Papa, G.L. Anthropogenic Soils: General Aspects and Features. *Ecocycles* **2015**, *1*, 3–8. [\[CrossRef\]](#)
- Davidson, E.; Trumbore, S.E.; Amundson, R. Soil Warming and Organic Carbon Content. *Nat. Cell Biol.* **2000**, *408*, 789–790. [\[CrossRef\]](#)
- Gholizadeh, A.; Saberioon, M.; Viscarra Rossel, R.A.; Boruvka, L.; Klement, A. Spectroscopic Measurements and Imaging of Soil Colour for Field Scale Estimation of Soil Organic Carbon. *Geoderma* **2020**, *357*, 113972. [\[CrossRef\]](#)
- Piao, S.; Fang, J.; Ciais, P.; Peylin, P.; Huang, Y.; Sitch, S.; Wang, T. The Carbon Balance of Terrestrial Ecosystems in China. *Nat. Cell Biol.* **2009**, *458*, 1009–1013. [\[CrossRef\]](#) [\[PubMed\]](#)
- Castaldi, F.; Chabrilat, S.; Chartin, C.; Genot, V.; Jones, A.; Van Wesemael, B. Estimation of Soil Organic Carbon in Arable Soil in Belgium and Luxembourg with the LUCAS Topsoil Database. *Eur. J. Soil Sci.* **2018**, *69*, 592–603. [\[CrossRef\]](#)
- Grover, S.; Butterly, C.R.; Wang, X.; Gleeson, D.B.; Macdonald, L.M.; Hall, D.; Tang, C. An Agricultural Practise with Climate and Food Security Benefits: “Claying” with Kaolinitic Clay Subsoil Decreased Soil Carbon Priming and Mineralisation in Sandy Cropping Soils. *Sci. Total Environ.* **2020**, *709*, 134488. [\[CrossRef\]](#)
- McKenzie, N.; Cresswell, H.P.; Ryan, P.J.; Grundy, M.J. Contemporary Land Resource Survey Requires Improvements in Direct Soil Measurement. *Commun. Soil Sci. Plant. Anal.* **2000**, *31*, 1553–1569. [\[CrossRef\]](#)
- Shepherd, K.D.; Walsh, M.G. Development of Reflectance Spectral Libraries for Characterization of Soil Properties. *Soil Sci. Soc. Am. J.* **2002**, *66*, 988–998. [\[CrossRef\]](#)
- Ben-Dor, E.; Banin, A. Near-Infrared Analysis as a Rapid Method to Simultaneously Evaluate Several Soil Properties. *Soil Sci. Soc. Am. J.* **1995**, *59*, 364–372. [\[CrossRef\]](#)
- Chang, C.-W.; Laird, D.A. Near-Infrared Reflectance Spectroscopic Analysis of Soil C and N. *Soil Sci.* **2002**, *167*, 110–116. [\[CrossRef\]](#)
- Morra, M.J.; Hall, M.H.; Freeborn, L.L. Carbon and Nitrogen Analysis of Soil Fractions Using Near-Infrared Reflectance Spectroscopy. *Soil Sci. Soc. Am. J.* **1991**, *55*, 288–291. [\[CrossRef\]](#)
- Angelopoulou, T.; Tziolas, N.; Balafoutis, A.; Zalidis, G.; Bochtis, D. Remote Sensing Techniques for Soil Organic Carbon Estimation: A Review. *Remote Sens.* **2019**, *11*, 676. [\[CrossRef\]](#)
- Hong, Y.; Chen, Y.; Yu, L.; Liu, Y.; Liu, Y.; Zhang, Y.; Liu, Y.; Cheng, H. Combining Fractional Order Derivative and Spectral Variable Selection for Organic Matter Estimation of Homogeneous Soil Samples by VIS–NIR Spectroscopy. *Remote Sens.* **2018**, *10*, 479. [\[CrossRef\]](#)
- Kühnel, A.; Bogner, C. In-Situ Prediction of Soil Organic Carbon by Vis-NIR Spectroscopy: An Efficient Use of Limited Field Data. *Eur. J. Soil Sci.* **2017**, *68*, 689–702. [\[CrossRef\]](#)
- Liu, Y.; Jiang, Q.; Fei, T.; Wang, J.; Shi, T.; Guo, K.; Li, X.; Chen, Y. Transferability of a Visible and Near-Infrared Model for Soil Organic Matter Estimation in Riparian Landscapes. *Remote Sens.* **2014**, *6*, 4305–4322. [\[CrossRef\]](#)
- Garrity, D.; Bindraban, P. *A Globally Distributed Soil Spectral Library, Visible Near Infrared Diffuse Reflectance Spectra*; The ICRAF/ISRIC Spectral Library; Soil-Plant Spectral Diagnostics Laboratory: Nairobi, Kenya, 2004.
- Orgiazzi, A.; Ballabio, C.; Panagos, P.; Jones, A.; Fernández-Ugalde, O. LUCAS Soil, the Largest Expandable Soil Dataset for Europe: A Review. *Eur. J. Soil Sci.* **2017**, *69*, 140–153. [\[CrossRef\]](#)
- Demattê, J.A.; Dotto, A.C.; Paiva, A.F.; Sato, M.V.; Dalmolin, R.S.; De Araújo, M.D.S.B.; Da Silva, E.B.; Nanni, M.R.; Caten, A.T.; Noronha, N.C.; et al. The Brazilian Soil Spectral Library (BSSL): A General View, Application and Challenges. *Geoderma* **2019**, *354*, 113793. [\[CrossRef\]](#)

22. Viscarra Rossel, R.A.; Webster, R. Predicting Soil Properties from the Australian Soil Visible-Near Infrared Spectroscopic Database. *Eur. J. Soil Sci.* **2012**, *63*, 848–860. [[CrossRef](#)]
23. Shi, Z.; Wang, Q.; Peng, J.; Ji, W.; Liu, H.; Li, X.; Viscarra Rossel, R.A. Development of a National VNIR Soil-Spectral Library for Soil Classification and Prediction of Organic Matter Concentrations. *Sci. China Earth Sci.* **2014**, *57*, 1671–1680. [[CrossRef](#)]
24. Stevens, A.; Nocita, M.; Tóth, G.; Montanarella, L.; Van Wesemael, B. Prediction of Soil Organic Carbon at the European Scale by Visible and Near InfraRed Reflectance Spectroscopy. *PLoS ONE* **2013**, *8*, e66409. [[CrossRef](#)] [[PubMed](#)]
25. Guerrero, C.; Wetterlind, J.; Stenberg, B.; Mouazen, A.M.; Gabarrón-Galeote, M.A.; Ruiz-Sinoga, 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](#)]
26. Jin, X.; Du, J.; Liu, H.; Wang, Z.; Song, K. Remote Estimation of Soil Organic Matter Content in the Sanjiang Plain, Northeast China: The Optimal Band Algorithm Versus the GRA-ANN Model. *Agric. Meteorol.* **2016**, *218*, 250–260. [[CrossRef](#)]
27. 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](#)]
28. Wetterlind, J.; Stenberg, B.; Söderström, M. Farm-Soil Mapping Using NIR-Technique for Increased Sample Point Density. In *Precision Agriculture 2007—Papers Presented at the 6th European Conference on Precision Agriculture*; Evangelical Christian Publishers Association (ECPA): Phoenix, AZ, USA, 2007; pp. 265–270.
29. Stenberg, B.; Wetterlind, J. Small Sized Local vs. Large Sized National Calibration Sets and Their Combination for Farm Scale Predictions by NIR. In *Geophysical Research Abstracts*; European Geosciences Union: Munich, Germany, 2009.
30. Wetterlind, J.; Stenberg, B.; Söderström, M. The Use of Near Infrared (NIR) Spectroscopy to Improve Soil Mapping at the Farm Scale. *Precis. Agric.* **2008**, *9*, 57–69. [[CrossRef](#)]
31. Mehmood, T.; Liland, K.H.; Snipen, L.; Sæbø, S. A Review of Variable Selection Methods in Partial Least Squares Regression. *Chemom. Intell. Lab. Syst.* **2012**, *118*, 62–69. [[CrossRef](#)]
32. Vohland, M.; Ludwig, M.; Thiele-Bruhn, S.; Ludwig, B. Quantification of Soil Properties with Hyperspectral Data: Selecting Spectral Variables with Different Methods to Improve Accuracies and Analyze Prediction Mechanisms. *Remote Sens.* **2017**, *9*, 1103. [[CrossRef](#)]
33. Vohland, M.; Ludwig, M.; Harbich, M.; Emmerling, C.; Thiele-Bruhn, S. Using Variable Selection and Wavelets to Exploit the Full Potential of Visible–Near Infrared Spectra for Predicting Soil Properties. *J. Near Infrared Spectrosc.* **2016**, *24*, 255–269. [[CrossRef](#)]
34. Chong, I.-G.; Jun, C.-H. Performance of Some Variable Selection Methods When Multicollinearity Is Present. *Chemom. Intell. Lab. Syst.* **2005**, *78*, 103–112. [[CrossRef](#)]
35. Jia, S.; Li, H.; Wang, Y.; Tong, R.; Li, Q. Recursive Variable Selection to Update Near-Infrared Spectroscopy Model for the Determination of Soil Nitrogen and Organic Carbon. *Geoderma* **2016**, *268*, 92–99. [[CrossRef](#)]
36. Galvão, R.K.H.; Araujo, M.C.U.; Frago, W.D.; Da Silva, E.C.; José, G.E.; Soares, S.F.C.; Paiva, H.M. A Variable Elimination Method to Improve the Parsimony of MLR Models Using the Successive Projections Algorithm. *Chemom. Intell. Lab. Syst.* **2008**, *92*, 83–91. [[CrossRef](#)]
37. Li, H.; Liang, Y.-Z.; Xu, Q.; Cao, D. Key Wavelengths Screening Using Competitive Adaptive Reweighted Sampling Method for Multivariate Calibration. *Anal. Chim. Acta* **2009**, *648*, 77–84. [[CrossRef](#)] [[PubMed](#)]
38. Leardi, R.; González, A.L. Genetic Algorithms Applied to Feature Selection in PLS Regression: How and When to Use Them. *Chemom. Intell. Lab. Syst.* **1998**, *41*, 195–207. [[CrossRef](#)]
39. Kalivas, J.H.; Roberts, N.; Sutter, J.M. Global Optimization by Simulated Annealing with Wavelength Selection for Ultraviolet-Visible Spectrophotometry. *Anal. Chem.* **1989**, *61*, 2024–2030. [[CrossRef](#)]
40. Li, H.-D.; Xu, Q.-S.; Liang, Y.-Z. Random Frog: An Efficient Reversible Jump Markov Chain Monte Carlo-Like Approach for Variable Selection with Applications to Gene Selection and Disease Classification. *Anal. Chim. Acta* **2012**, *740*, 20–26. [[CrossRef](#)]
41. Zhang, Y.; Li, M.; Zheng, L.; Qin, Q.; Lee, W.S. Spectral Features Extraction for Estimation of Soil Total Nitrogen Content Based on Modified Ant Colony Optimization Algorithm. *Geoderma* **2019**, *333*, 23–34. [[CrossRef](#)]

42. Vohland, M.; Ludwig, M.; Thiele-Bruhn, S.; Ludwig, B. Determination of Soil Properties with Visible to Near- and Mid-Infrared Spectroscopy: Effects of Spectral Variable Selection. *Geoderma* **2014**, *223–225*, 88–96. [\[CrossRef\]](#)
43. Yao, X.; Yang, W.; Li, M.; Zhou, P.; Chen, Y.; Hao, Z.; Liu, Z. Prediction of Total Nitrogen in Soil Based on Random Frog Leaping Wavelet Neural Network. *IFAC Pap.* **2018**, *51*, 660–665. [\[CrossRef\]](#)
44. Hu, M.; Dong, Q.; Liu, B.-L.; Opara, U.L.; Chen, L. Estimating Blueberry Mechanical Properties Based on Random Frog Selected Hyperspectral Data. *Postharvest Biol. Technol.* **2015**, *106*, 1–10. [\[CrossRef\]](#)
45. Li, X.; Sun, C.; Luo, L.; He, Y. Determination of Tea Polyphenols Content by Infrared Spectroscopy Coupled with IPLS and Random Frog Techniques. *Comput. Electron. Agric.* **2015**, *112*, 28–35. [\[CrossRef\]](#)
46. Gholizadeh, A.; Borůvka, L.; Saberioon, M.; Kozák, J.; Vašát, R.; Němeček, K. Comparing Different Data Preprocessing Methods for Monitoring Soil Heavy Metals Based on Soil Spectral Features. *Soil Water Res.* **2016**, *10*, 218–227. [\[CrossRef\]](#)
47. Vašát, R.; Kodešová, R.; Klement, A.; Borůvka, L. Simple but Efficient Signal Pre-Processing in Soil Organic Carbon Spectroscopic Estimation. *Geoderma* **2017**, *298*, 46–53. [\[CrossRef\]](#)
48. Gholizadeh, A.; Carmon, N.; Klement, A.; Ben-Dor, E.; Boruvka, L. Agricultural Soil Spectral Response and Properties Assessment: Effects of Measurement Protocol and Data Mining Technique. *Remote Sens.* **2017**, *9*, 1078. [\[CrossRef\]](#)
49. Rinnan, Å.; Berg, F.V.D.; Engelsen, S.B. Review of the Most Common Pre-Processing Techniques for Near-Infrared Spectra. *Trac. Trends Anal. Chem.* **2009**, *28*, 1201–1222. [\[CrossRef\]](#)
50. Gholizadeh, A.; Žižala, D.; Saberioon, M.; Borůvka, L. Soil Organic Carbon and Texture Retrieving and Mapping Using Proximal, Airborne and Sentinel-2 Spectral Imaging. *Remote Sens. Environ.* **2018**, *218*, 89–103. [\[CrossRef\]](#)
51. Echambadi, R.; Hess, J.D. Mean-Centering Does Not Alleviate Collinearity Problems in Moderated Multiple Regression Models. *Mark. Sci.* **2007**, *26*, 438–445. [\[CrossRef\]](#)
52. Wu, Z.; Wang, B.; Huang, J.; An, Z.; Jiang, P.; Chen, Y.; Liu, Y. Estimating Soil Organic Carbon Density in Plains Using Landscape Metric-Based Regression Kriging Model. *Soil Tillage Res.* **2019**, *195*, 104381. [\[CrossRef\]](#)
53. Shi, T.; Chen, Y.; Liu, H.; Wang, J.; Wu, G. Soil Organic Carbon Content Estimation with Laboratory-Based Visible–Near-Infrared Reflectance Spectroscopy: Feature Selection. *Appl. Spectrosc.* **2014**, *68*, 831–837. [\[CrossRef\]](#)
54. Rossel, R.V.; Behrens, T. Using Data Mining to Model and Interpret Soil Diffuse Reflectance Spectra. *Geoderma* **2010**, *158*, 46–54. [\[CrossRef\]](#)
55. Shi, Z.; Ji, W.; Viscarra Rossel, R.A.; Chen, S.; Zhou, Y. Prediction of Soil Organic Matter Using a Spatially Constrained Local Partial Least Squares Regression and the Chinese Vis-NIR Spectral Library. *Eur. J. Soil Sci.* **2015**, *66*, 679–687. [\[CrossRef\]](#)
56. Pansu, M.; Gautheyrou, J. *Handbook of Soil Analysis: Mineralogical Organic and Inorganic Methods*; Springer: Berlin/Heidelberg, Germany, 2006.
57. Wold, S.; Martens, H.; Wold, H. *The Multivariate Calibration Problem in Chemistry Solved by the PLS Method*; Springer: Berlin/Heidelberg, Germany, 1983; pp. 286–293.
58. 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. [\[CrossRef\]](#)
59. Liu, Y.; Liu, Y.; Chen, Y.; Zhang, Y.; Shi, T.; Wu, G.; Hong, Y.; Fei, T. The Influence of Spectral Pretreatment on the Selection of Representative Calibration Samples for Soil Organic Matter Estimation Using Vis-NIR Reflectance Spectroscopy. *Remote Sens.* **2019**, *11*, 450. [\[CrossRef\]](#)
60. Cai, S.; Xia, X. The Wetland Resource of the Sihua Area and Its Exploitation. *J. Resour. Environ. Yangtze Val.* **1993**, *2*, 137–141.
61. Wang, X.; Chen, Y.; Guo, L.; Liu, L. Construction of the Calibration Set through Multivariate Analysis in Visible and Near-Infrared Prediction Model for Estimating Soil Organic Matter. *Remote Sens.* **2017**, *9*, 201. [\[CrossRef\]](#)
62. Liu, Y.; Chen, Y. Estimation of Total Iron Content in Floodplain Soils Using VNIR Spectroscopy—A Case Study in the Le'an River Floodplain, China. *Int. J. Remote Sens.* **2012**, *33*, 5954–5972. [\[CrossRef\]](#)
63. Viscarra Rossel, R.A.; Minasny, B.; Roudier, P.; McBratney, A.B. Colour Space Models for Soil Science. *Geoderma* **2006**, *133*, 320–337. [\[CrossRef\]](#)
64. Lacerda, M.P.C.; Demattê, J.A.; Sato, M.V.; Fongaro, C.T.; Gallo, B.; Souza, A.B. Tropical Texture Determination by Proximal Sensing Using a Regional Spectral Library and Its Relationship with Soil Classification. *Remote Sens.* **2016**, *8*, 701. [\[CrossRef\]](#)

65. Ladoni, M.; Bahrami, H.A.; Alavipanah, S.K.; Noroozi, A.A. Estimating Soil Organic Carbon from Soil Reflectance: A Review. *Precis. Agric.* **2009**, *11*, 82–99. [[CrossRef](#)]
66. Stenberg, B.; Viscarra Rossel, R.A.; Mouazen, A.M.; Wetterlind, J. Visible and Near Infrared Spectroscopy in Soil Science. In *Advances in Agronomy*; Academic Press: Cambridge, MA, USA, 2010; Volume 107, pp. 163–215.

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



© 2020 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 (<http://creativecommons.org/licenses/by/4.0/>).