



Article

Application of Spectrally Derived Soil Type as Ancillary Data to Improve the Estimation of Soil Organic Carbon by Using the Chinese Soil Vis-NIR Spectral Library

Yi Liu ^{1,2} , Zhou Shi ³ , Ganlin Zhang ², Yiyun Chen ^{1,3,4,*} , Shuo Li ⁵, Yongshen Hong ¹, Tiezhu Shi ⁶, Junjie Wang ⁶ and Yaolin Liu ^{1,*}

¹ School of Resource and Environment Science, Wuhan University, 129 Luoyu Road, Wuhan 430079, China; liuyi2010@whu.edu.cn (Y.L.); hys@whu.edu.cn (Y.H.)

² State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, Nanjing 210008, China; glzhang@issas.ac.cn

³ Institute of Applied Remote Sensing and Information Technology, College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China; shizhou@zju.edu.cn

⁴ Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, China

⁵ The college of Urban & Environmental Science, Central China Normal University, 152 Luoyu Road, Wuhan 430079, China; shuoguo@gmail.com

⁶ Key Laboratory for Geo-Environmental Monitoring of Coastal Zone of the National Administration of Surveying, Mapping and GeoInformation & Shenzhen Key Laboratory of Spatial Smart Sensing and Services, Shenzhen University, Shenzhen 518060, China; tiezhushi@szu.edu.cn (T.S.); wjlight@whu.edu.cn (J.W.)

* Correspondence: chenyy@whu.edu.cn (Y.C.); liuyaolin2010@163.com (Y.L.)

Received: 18 September 2018; Accepted: 3 November 2018; Published: 6 November 2018



Abstract: Ancillary data, such as soil type, may improve the visible and near-infrared (vis-NIR) estimation of soil organic carbon (SOC); however, they require data collection or expert knowledge. The application of a national soil spectral library to local SOC estimations usually requires soil type information, because the relationships between vis-NIR spectra and SOC from different populations may vary. Using 515 samples of five soil types (genetic soil classification of China, GSCC) from the Chinese soil spectral library (CSSL), we compared three strategies in the vis-NIR estimation of SOC. Different regression models were calibrated using the entire dataset (Strategy I, without using soil type as ancillary data) and the subsets stratified by soil type from CSSL as ancillary data (strategies II and III). In Strategy II, the subsets were stratified by soil type from the CSSL for validation. In Strategy III, the subsets were stratified by spectrally derived soil type for validation. The results showed that 86.72% of the samples were successfully discriminated for the soil types by using the vis-NIR spectra. The coefficients of determination in the prediction (R_p^2) of SOC estimation by strategies I, II, and III were 0.74, 0.83, and 0.82, respectively. The stratified calibration strategies (strategies II and III) improved the vis-NIR estimation of SOC. The misclassification of the soil type in the application of Strategy III slightly affected the SOC estimations. Nevertheless, this strategy is inexpensive and beneficial when expert knowledge on soil classification is lacking. We concluded that vis-NIR spectroscopy could be applied to distinguish some soil types in terms of GSCC, which further provided essential and easily accessible ancillary data for the application of stratified calibration strategies in the vis-NIR estimation of SOC.

Keywords: soil spectral library; vis-NIR spectroscopy; soil organic carbon; soil type

1. Introduction

The content of soil organic carbon (SOC; 1550 Gt) is higher than that of the combined carbon from global vegetation (420–620 Gt) and the atmosphere (760 Gt) [1,2]. Even though a small proportion of SOC is transformed into atmospheric carbon as greenhouse gases, its potential influence on the global climate is substantial [3]. It is well recognized that SOC is important for sustaining soil quality and food production, and inappropriate land-use management practices might cause the loss of SOC [4,5]. Due to the critical role of SOC in food production and climate regulation, the demand for monitoring the spatial and temporal variation of SOC is increasing [6]. The conventional laboratory analysis of SOC such as combustion or chromate oxidation [7,8] is expensive and time-consuming [9]. Thus, techniques for the rapid and inexpensive measurement of SOC should be developed.

Visible and near-infrared (vis-NIR) diffuse reflectance spectroscopy has been rapidly developed as an alternative to a conventional laboratory analysis of soil properties with an acceptable level of accuracy [10,11]. Vis-NIR spectroscopy has many advantages; it requires less sample preparation; is inexpensive, rapid, and non-destructive; it can be used for the simultaneous estimation of various soil properties; it needs no or less chemical reagents [12]. Furthermore, it can be obtained at proximal and remote sensing platforms, such as in situ and airborne sensors, to assess soil properties [3,13,14].

Constructing vis-NIR spectroscopy models in a specific geographical region requires a soil library to relate soil spectra to soil property through multivariate regressions. Such soil libraries must encompass a wide variation in soil property [4]. Over the past decade, soil libraries have been built on various scales ranging from field or local to national, continental, or global scales [15–18]. Recently, the challenge has shifted from soil library building to its application. How to properly employ a large soil library to estimate soil properties has become a hot topic. The most commonly utilized strategies to facilitate the application of large libraries include powerful regression approaches [19–21], optimal spectral transformations [22–24], representative calibration sample selection [10,17,25–28], ancillary data integration [4,14,15,29–33], subset spiking, and extra weighting [34–37]. A large soil library usually consists of various samples in terms of geographical origins, minerals, parent materials, environmental conditions, and land-use types. Soil type can be a comprehensive indicator of different soil populations, because the soil classification system considers multiple factors. Different soil types can vary from one another in terms of the relationship between vis-NIR spectra and SOC [38]. Some researchers have used soil type to stratify soil libraries, and suggested that soil type may help improve soil property estimation through vis-NIR spectroscopy [14,32,33].

Despite the advantages of ancillary data—such as soil type—in soil property estimation through vis-NIR spectroscopy, data collection imposes an extra cost burden and requires expert knowledge. Thus, easily accessible ancillary data have been preferred. Furthermore, vis-NIR spectra can be a good predictor of soil types [39,40]. However, whether spectrally derived soil type can improve the vis-NIR estimation of SOC requires further investigation. The potential of vis-NIR spectroscopy for the provision of soil type data to estimate SOC should also be explored.

This study explored the application of spectrally derived and actual soil types as ancillary data to improve SOC estimation through vis-NIR spectroscopy and the Chinese soil library. Specifically, this study investigated the following. (i) We discriminated soil type through partial least squares discriminant analysis (PLS-DA) and examined its classification accuracy. (ii) We calibrated partial least squares regression (PLSR) models by using the entire dataset (Strategy I) and the subsets stratified by soil type from the Chinese soil spectral library (CSSL, strategies II and III). In Strategy II, we stratified the subsets by soil type from CSSL for validation. In Strategy III, we stratified the subsets by spectrally derived soil type for validation.

2. Materials and Methods

2.1. Sample Collection

The CSSL (CSSL-2014) comprised 1581 samples from 14 provinces out of China's 34 provinces (autonomous regions, municipalities, special administrative regions) with multiple land-use and land-cover types (Supplementary Materials Figure S1). Most samples were from cultivated land with intensive farming. Shi et al. [10,41] also described the spatial distribution of the samples in detail. CSSL represents 16 soil types based on the genetic soil classification of China (GSCC). The GSCC is different from the United States (US) Soil Taxonomy System and World Reference Base for Soil Resources (WRB). We did not transform the soil types of GSCC to those of the other two classification systems because no accurate transformations among the three classification systems exist. Their most possible soil types in WRB classification are provided in Table 1 [42,43].

Table 1. Sample divisions and descriptive statistics of soil organic carbon (SOC) ($\text{g}\cdot\text{kg}^{-1}$).

Soil Type ^a	WRB ^b	Sample Set	Count	Min	Max	Mean	SD ^c	Skewness	Kurtosis	CV ^d
Coastal solonchaks	Solonchaks	All	114	2.15	18.27	7.23	2.63	0.75	2.02	0.36
		Calibration	86	2.15	18.27	7.26	2.72	0.87	2.32	0.37
		Validation	28	2.61	12.06	7.13	2.35	0.07	−0.45	0.33
Meadow soils	Cambisols	All	52	9.80	27.26	17.67	3.78	0.08	−0.03	0.21
		Calibration	39	9.80	27.26	17.61	3.75	0.07	0.07	0.21
		Validation	13	10.73	25.58	17.85	4.01	0.08	−0.3	0.22
Chernozems	Chernozems	All	138	6.38	25.29	14.73	3.27	−0.02	−0.02	0.22
		Calibration	104	6.38	25.29	14.76	3.31	0.04	0.09	0.22
		Validation	34	7.25	20.48	14.64	3.17	−0.26	−0.54	0.22
Black soils	Phaeozems	All	104	6.96	33.99	16.61	4.38	1.01	2.68	0.26
		Calibration	78	6.96	33.99	16.57	4.38	0.99	2.88	0.26
		Validation	26	9.11	30.63	16.75	4.46	1.05	2.09	0.27
Purplish soils	Cambisols	All	107	0.96	25.17	11.75	5.44	0.05	−0.86	0.46
		Calibration	80	0.96	25.17	11.76	5.44	0.06	−0.83	0.46
		Validation	27	1.52	22.45	11.73	5.57	0.02	−0.94	0.48
Total		All	515	0.96	33.99	13.13	5.39	0.1	−0.17	0.41
		Calibration	387	0.96	33.99	13.12	5.38	0.1	−0.15	0.41
		Validation	128	1.52	30.63	13.14	5.44	0.11	−0.21	0.41

^a Soil type in the table refer to the genetic soil classification of China (National Soil Survey Office, 1996). ^b World Reference Base for Soil Resources (WRB) (IUSS Working Group WRB, 2007). ^c SD denotes standard deviation. ^d CV denotes coefficient of variation.

In this study, 515 samples from five soil types were considered, and they were diverse in terms of SOC. For example, some soil types had low (coastal solonchaks) or high (meadow soils) SOC concentration, whereas some had moderate (purplish soils) SOC concentration. The five soil types comprised a moderate number of samples (52–138 samples). Some soil types, such as alluvial soils ($n = 11$) and paddy soils ($n = 552$), containing extremely large or small number of samples in the CSSL, were not selected. The number of samples for each soil type in previous similar studies was as follows: 8–3928 [40], 2–66 [39], 82–2077 [14], 184–367 [33], and 26–75 [32]. Therefore, the number of samples for each type was reasonable in this study. The selected types included coastal solonchaks, meadow soils, chernozems, black soils, and purplish soils (Table 1). The other soil types are shown in the Supplementary Materials.

2.2. Spectral Measurement and Chemical Analysis

Topsoil samples (0–20 cm) were taken to the laboratory, air-dried, and ground to pass a 2-mm sieve before spectral measurement. Afterward, SOC analysis was performed. An ASD FieldSpec ProFR vis-NIR spectrometer (Analytical Spectral Devices, Boulder, CO, USA) with a spectral range of 350–2500 nm was used [44]. Spectral measurement was conducted in a dark room with a halogen lamp as a light source, which was positioned 7 cm away from the soil samples with a 30° zenith angle. The soil samples were placed in a 10 cm-diameter Petri dish with a thickness of 1.5 cm. The fiber probe

was installed 15 cm above the soil samples with a view angle of 25°. A Spectralon® panel with 99% reflectance was utilized to calibrate the spectrometer before measurement. Each sample was scanned 10 times and averaged [41]. SOC was determined with a potassium dichromate volumetric external heating method in accordance with Chinese standards (specification of soil test, SL237-1999) [45].

Several spectral pretreatments were used to reduce noise and enhance the spectral features, because raw spectra might be influenced by the working status of a spectroradiometer, experimental conditions, particle sizes, and surface roughness [3,46]. The two edges with low signal-to-noise ratios, namely, 350–399 nm and 2451–2500 nm, were first removed [47]. The six commonly used pretreatments or their combinations, namely, Savitzky–Golay smoothing [48], logarithmic function ($\log 1/R$) [49], first derivative [50], standard normal variate [51], multiplicative scatter correction [52], and optimal pretreatments, were then tested and utilized (Supplementary Materials Table S1). The mean center was applied to improve the numerical stability of some methods (e.g., PLSR) [53–55].

2.3. Model Calibration and Validation

PLSR was used to correlate the soil spectral data with SOC [56–58], and leave-one-out cross-validation was utilized to determine the optimal number of latent variables [59]. To assess the predictive ability of the models, we applied several commonly used indicators, namely, root mean square error of cross validation ($RMSE_{cv}$), coefficient of determination in cross validation (R_{cv}^2), root mean square error of prediction ($RMSEP$), coefficient of determination in prediction (R_p^2), and residual predictive deviation (RPD), as expressed in the equations in the Supplementary Materials [13]. R_p^2 and RPD could be equivalent [60,61]. Both indicators were reported to ensure that our study could be used for comparison by other researchers when they used either or both of them. However, the discussion was based on R_p^2 .

2.3.1. Model Calibration

The division of 25%/75% for validation and calibration was based on the ascending order of SOC concentration. Subsequently, the samples of the validation set were selected at intervals of three samples. Such division rather than a random selection was to ensure that the validation samples were evenly distributed in the range of the SOC concentration and covered the SOC diversity of expected future samples [49]. Moreover, this division was commonly adopted in previous studies [3,44,62]. This division is characterized by some limitations, including the need for previous information regarding the SOC concentration and the seemingly arbitrary or empirical choice of the first validation sample and the number of the interval.

Strategy I utilized the entire calibration samples ($515 \times 75\% = 387$ samples) to build a PLSR model for estimating SOC. For strategies II and III, 387 calibration samples were stratified into five subsets by soil type to build five separate PLSR models for estimating SOC.

2.3.2. Model Validation

Strategy I utilized all of the validation samples ($515 \times 25\% = 128$ samples) to validate the built PLSR model. For Strategy II, 128 validation samples were stratified by soil type and then allocated to the respectively built PLSR models. For Strategy III, the soil type of the 128 validation samples were assumed to be unknown and needed to be derived by the PLS-DA model by using their spectra. For Strategy III, the 128 validation samples were stratified by spectrally derived soil type and then allocated to the respective PLSR models.

The soil type of the validation samples was discriminated by PLS-DA before Strategy III was applied. PLS-DA, which was developed based on PLSR, directly relates the variables in the spectral data to soil types [63]. The calibration samples ($n = 387$) were utilized to build a PLS-DA model, and the vis-NIR spectra of the validation samples ($n = 128$) served as the inputs of the PLS-DA model. The soil type of the validation samples could then be discriminated. The agreement rate, which is the

proportion of the samples correctly predicted in the class, was used to evaluate the performance of the PLS-DA model.

3. Result

3.1. Descriptive Statistics

SOC concentrations varied from $0.96 \text{ g}\cdot\text{kg}^{-1}$ to $33.99 \text{ g}\cdot\text{kg}^{-1}$ with a mean of $13.13 \text{ g}\cdot\text{kg}^{-1}$ (Table 1 and Figure 1). The coefficient of variation (CV) was between 0.21 (meadow soils) and 0.48 (purplish soils). The SOC was $<12 \text{ g}\cdot\text{kg}^{-1}$ in most of the coastal solonchak samples, but $>12 \text{ g}\cdot\text{kg}^{-1}$ in most of the meadow soils samples. The three other soil types did not exhibit evident SOC boundaries. The kurtosis for coastal solonchak, black soil, and total samples was >2 , indicating that a mass of the samples concentrated around the center. The distribution of meadow soils (skewness = 0.08, kurtosis = -0.03) was close to normal distribution. Some black soil and coastal solonchak samples were deemed as outliers based on the boxplot. However, we still retained these samples in our models, because our data and aims were on a countrywide scale, and encountering samples with a high SOC concentration was expected.

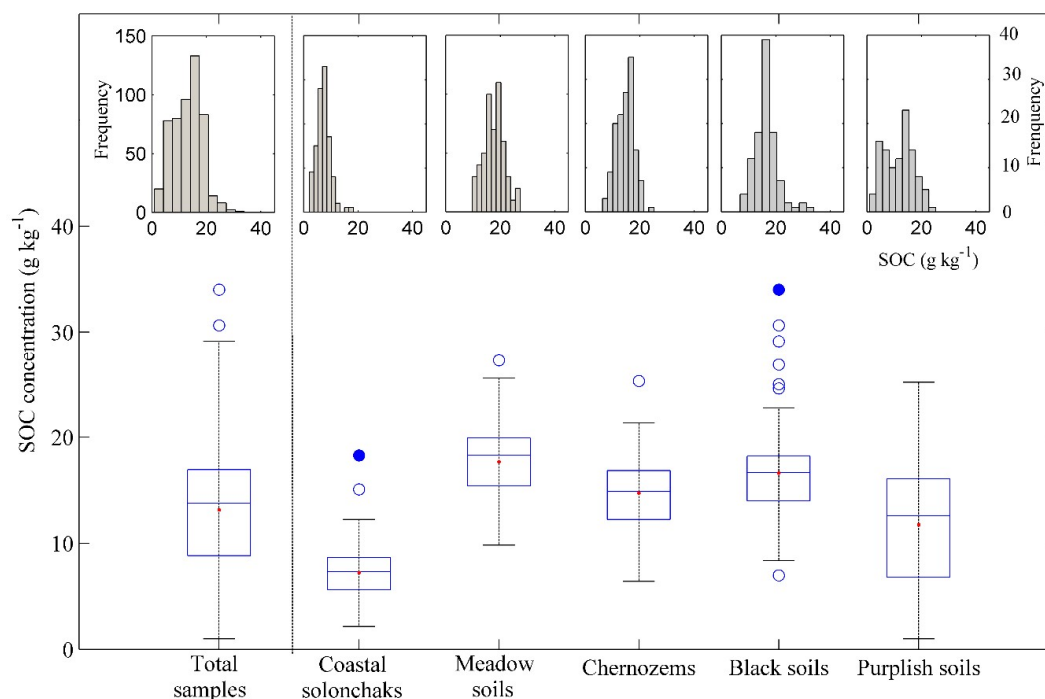


Figure 1. Boxplot and histogram of soil organic carbon concentration for five soil types from the Chinese soil spectral library. Redpoint (·), blue line, hollow circle (○), blue solid circle (●), and blue box denote the mean value, median value, outliers, extreme outliers, and interquartile range, respectively.

The statistical indicators of the calibration and validation set were similar. However, the validation of the coastal solonchak samples was closely distributed to the normal (skewness = 0.07, kurtosis = -0.45), which was quite different from the corresponding calibration samples. This finding was observed because we assigned an outlier with a high SOC to the calibration set. However, the other statistical indicators, minimum and CV, were similar. In the other soil types, we allocated an outlier of black soil samples into the validation set, and the validation set remained similar to the calibration set. In summary, our separation of the samples allowed the validation samples to cover the variation in the SOC of the soil library.

Figure 2 shows the average reflectance and SOC concentration of the soil samples from each soil type. The spectra showed three prominent absorption peaks at 1420 nm, 1920 nm, and 2210 nm;

the first two were mainly caused by the hydroxyl group (OH) of free water, and the last one was due to the Al–OH lattice structure in clay minerals [3]. The purplish soils and coastal solonchaks had lower SOC concentrations, but higher reflectance than the meadow soils, chernozems, and black soils. The spectral curves of meadow soils, chernozems, and black soils were close to one another and had overlapped at some bands, because their mean SOC concentrations were similar. Different soil types revealed diverse curves in shapes and SOC concentration. Therefore, including the soil type variable into SOC estimation can improve the estimation accuracy.

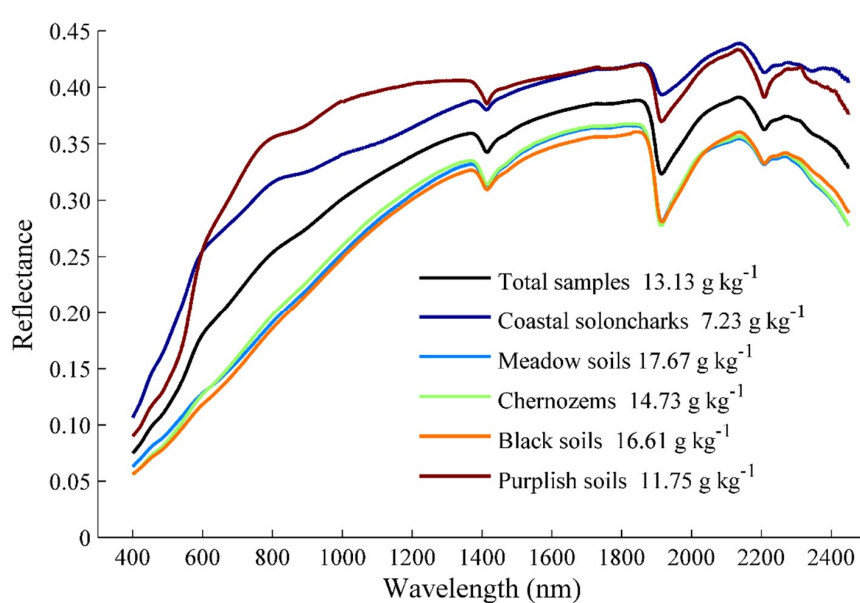


Figure 2. Mean reflectance of soil samples from five soil types in Chinese soil spectral library. The mean value of soil organic carbon concentration (SOC) of each soil type is marked.

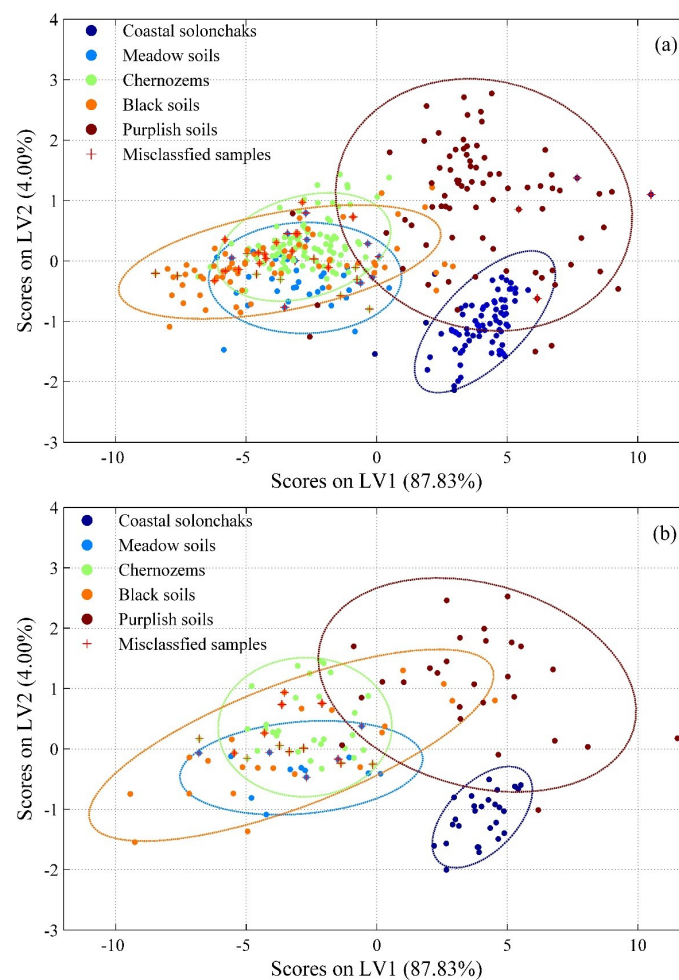
3.2. Discriminating Soil Type through Vis-NIR Spectroscopy

In the calibration set, 89.9% of the samples were correctly assigned (Table 2). Coastal solonchaks could be well distinguished from the others, with two out of 86 samples misclassified to purplish soils because of the similarities in their reflectance curves of these two soil types (Figure 2). Coastal solonchaks scattered away from other soil types except purplish soils, which was possible because its samples were collected from a concentrated location (Figure 3a). Approximately 97.50% of purplish soil samples were correctly discriminated, and only four were misclassified to black soils and one was misclassified to coastal solonchaks, because most of the purplish soils were laid far from other soil types, and only a few were found within the overlapping area (Figure 3a). Meadow soils, chernozems, and black soils were close and overlapping with one another (Figures 2 and 4a), and only 74.36–90.38% of their samples were correctly classified (Table 2).

In the validation set, all of the soil types except meadow soils were well distinguished, and the agreement rate was over 79% (Table 3). All of the coastal solonchaks and purplish soil samples were correctly discriminated. Meadow soils obtained the poorest accuracy, because their spectral characteristic features are similar to those of chernozems and black soils (Figures 2 and 3b). Nevertheless, the results for chernozems and black soils were acceptable. Similar to the results of the calibration set (89.92%), the overall agreement rate was 86.72% for the validation set.

Table 2. Confusion matrix of soil type prediction using partial least squares discriminant analysis (PLS-DA) and Chinese soil spectral library (CSSL).

		Actual Soil Order					Agreement Rate (%)
		Coastal Solonchaks	Meadow Soils	Chernozems	Black Soils	Purplish Soils	
Predicted soil type	Calibration set	Coastal solonchaks	84	0	0	0	97.67
		Meadow soils	0	29	9	5	74.36
		Chernozems	0	10	94	10	90.38
		Black soils	0	0	1	63	80.77
		Purplish soils	2	0	0	0	97.50
	Overall agreement rate (%)						89.92
Validation set	Validation set	Coastal solonchaks	28	0	0	0	100.00
		Meadow soils	0	8	6	1	61.54
		Chernozems	0	4	27	4	79.41
		Black soils	0	1	1	21	80.77
		Purplish soils	0	0	0	0	100.00
	Overall agreement rate (%)						86.72

**Figure 3.** Scatter diagram of scores on latent variable 2 (LV2) plotted against latent variable 1 (LV1) for calibration (a) and validation (b) in partial least squares discriminant analysis (PLS-DA) models. The samples are projected onto a plane defined by two latent variables. The ellipse is the 90% confidence ellipse for each soil type.

3.3. Estimation Accuracy of SOC Models Using Different Stratification Strategies

When the entire dataset was used to estimate SOC (Strategy I), the model performance of the five soil types is shown in Table 3 and Figure 4. The vis-NIR models tend to underestimate high SOC values and overestimate low SOC when the slope of the regression lines was generally <1 [14]. The poor model accuracy for the coastal solonchaks samples ($R_p^2 = 0.12$) was partly because of their low SOC concentration and right-tailed distribution (skewness = 2.68). The R_p^2 of meadow soils was only 0.47, because the small number of its samples cannot fully represent the relationship between SOC and the vis-NIR spectra of this soil type. Black soils exhibited an R_p^2 value of 0.46 because its few samples showed a high SOC concentration that resulted in a tail (skewness = 2.68, Figure 2). The R_p^2 of chernozems and purplish soils was above 0.6 because their mean SOC concentration was similar to that of the entire dataset. The overall R_p^2 (0.74) was higher than the R_p^2 of each soil type (0.12–0.72) because of the inner design and calculation of this statistical indicator (Supplementary Materials, Equation (4)). In summary, the model performed poorly in estimating SOC using the entire dataset for some soil types, especially coastal solonchaks.

Stratified calibration strategies improved the model performance in terms of cross-validation: the overall R_{cv}^2 increased from 0.62 to 0.75 (Table 3). For further validations, two strategies (strategies II and III) were proposed regarding the availability of soil type information. The soil type information of the validation samples that were used in Strategy II was obtained from CSSL, whereas that in Strategy III was derived through vis-NIR spectra.

The soil types of the validation samples were known (Strategy II), and the SOC of all of the soil types except coastal solonchaks were well estimated with $R_p^2 \geq 0.66$ (Table 3 and Figure 5). When the soil type of the validation samples was derived through vis-NIR spectroscopy (Strategy III), similar results were observed after stratification (Table 3).

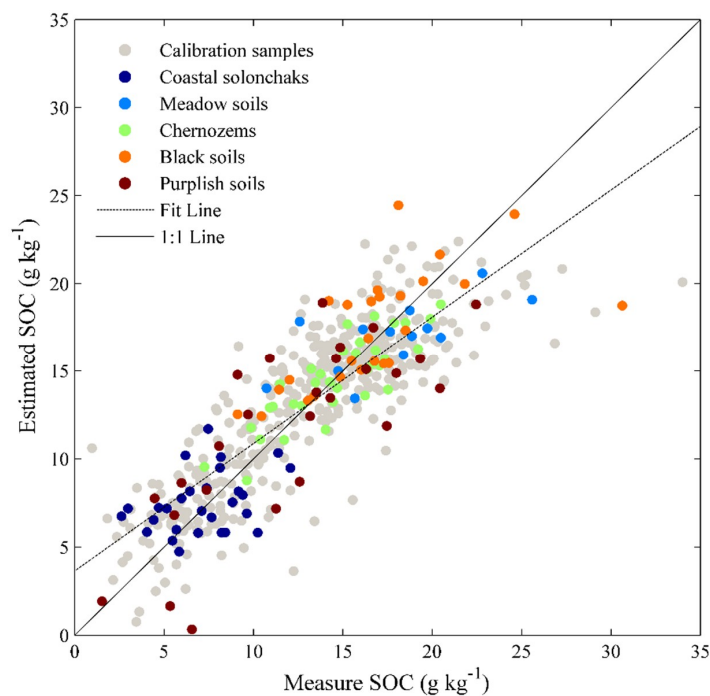
Compared with those in Strategy I, the SOC estimation was more accurate in strategies II and III, and the overall R_p^2 increased from 0.74 to 0.83 (Strategy II) and 0.82 (Strategy III) when the validation samples were stratified by soil type. The coastal solonchaks, which were poorly estimated ($R_p^2 = 0.12$) without stratification, had an acceptable accuracy with $R_p^2 = 0.51$ when the samples were stratified by soil type. The R_p^2 of meadow soils and black soils greatly increased from approximately 0.46–0.47 to 0.67–0.73. For chernozems and purplish soils, which were well estimated by Strategy I, stratification slightly improved the performance and changed their SOC estimation models because the subsets of chernozems and purplish soils resembled the entire dataset based on statistical indicators (Figure 2). By contrast, the subsets of coastal solonchaks, meadow soils, and black soils differed greatly from the entire dataset and were improved after stratification (Table 1). In summary, stratifying the soil library by soil type, including spectrally-derived soil type, enhanced the quality of vis-NIR models.

Comparison of the different methods of obtaining soil type (strategies II and III) revealed that the SOC estimation models stratified by spectrally-derived soil type (Strategy III) were slightly less robust, and the overall R_p^2 slightly decreased from 0.83 to 0.82. For coastal solonchaks and purplish soils, the two strategies produced the same result because of a 100% agreement rate (Table 3). For meadow soils and chernozems, the stratification by spectrally derived soil type achieved a slightly less accurate model ($R_p^2 = 0.67$ and 0.73) than the stratification by actual soil type ($R_p^2 = 0.73$ and 0.77). However, a large number of samples (38.46% and 20.49%) were misclassified into other groups. For black soils, a slight improvement was observed, and the R_p^2 increased from 0.70 to 0.72. In summary, the effect of misclassification was limited, and will be further discussed in the next section.

Table 3. Summary statistics for the estimation models of soil organic carbon (SOC) by partial least squares regression (PLSR).

Soil Type	Calibration			Validation			LVs
	R^2_{cv}	RMSE $_{cv}$	SD	R^2_p	RMSEP	RPD	
Entire dataset (not stratified)							
Coastal solonchaks	0.48	2.01	1.73	0.12	2.38	0.99	12
Meadow soils	0.37	3.01	2.02	0.47	3.03	1.32	12
Chernozems	0.41	2.53	2.48	0.72	1.66	1.91	12
Black soils	0.17	4.06	3.34	0.46	3.28	1.36	12
Purplish soils	0.37	4.59	5.75	0.63	3.66	1.52	12
Overall	0.62	3.35	4.99	0.74	2.80	1.94	12
Stratified by soil type							
Coastal solonchaks	0.56	1.80	1.76	0.51	1.63	1.44	3
Meadow soils	0.67	2.18	2.86	0.73	2.10	1.91	7
Chernozems	0.73	1.73	3.29	0.77	1.59	2.00	11
Black soils	0.43	3.40	4.18	0.70	2.49	1.79	16
Purplish soils	0.53	3.70	4.46	0.66	3.18	1.75	4
Overall	0.75	2.27	5.22	0.83	2.26	2.41	-
Stratified by soil type							
Coastal solonchaks	0.56	1.80	1.76	0.51	1.63	1.44	3
Meadow soils	0.67	2.18	2.47	0.67	2.37	1.69	7
Chernozems	0.73	1.73	3.17	0.73	1.71	1.85	11
Black soils	0.43	3.40	4.45	0.72	2.45	1.82	16
Purplish soils	0.53	3.70	4.46	0.66	3.18	1.75	4
Overall	0.75	2.27	5.22	0.82	2.30	2.37	-

Note: R^2_{cv} denotes the coefficient of determination in cross-validation, RMSE_{cv} denotes root-mean-square error of cross-validation, RMSEP denotes root mean square error of prediction, R^2_p denotes coefficient of determination in prediction, RPD denotes residual predictive deviation, and LV denotes latent variable, SD denotes the standard deviation of estimated SOC concentration.

**Figure 4.** Estimated versus measured soil organic carbon (SOC) plots of spectroscopy models with the entire dataset (Strategy I).

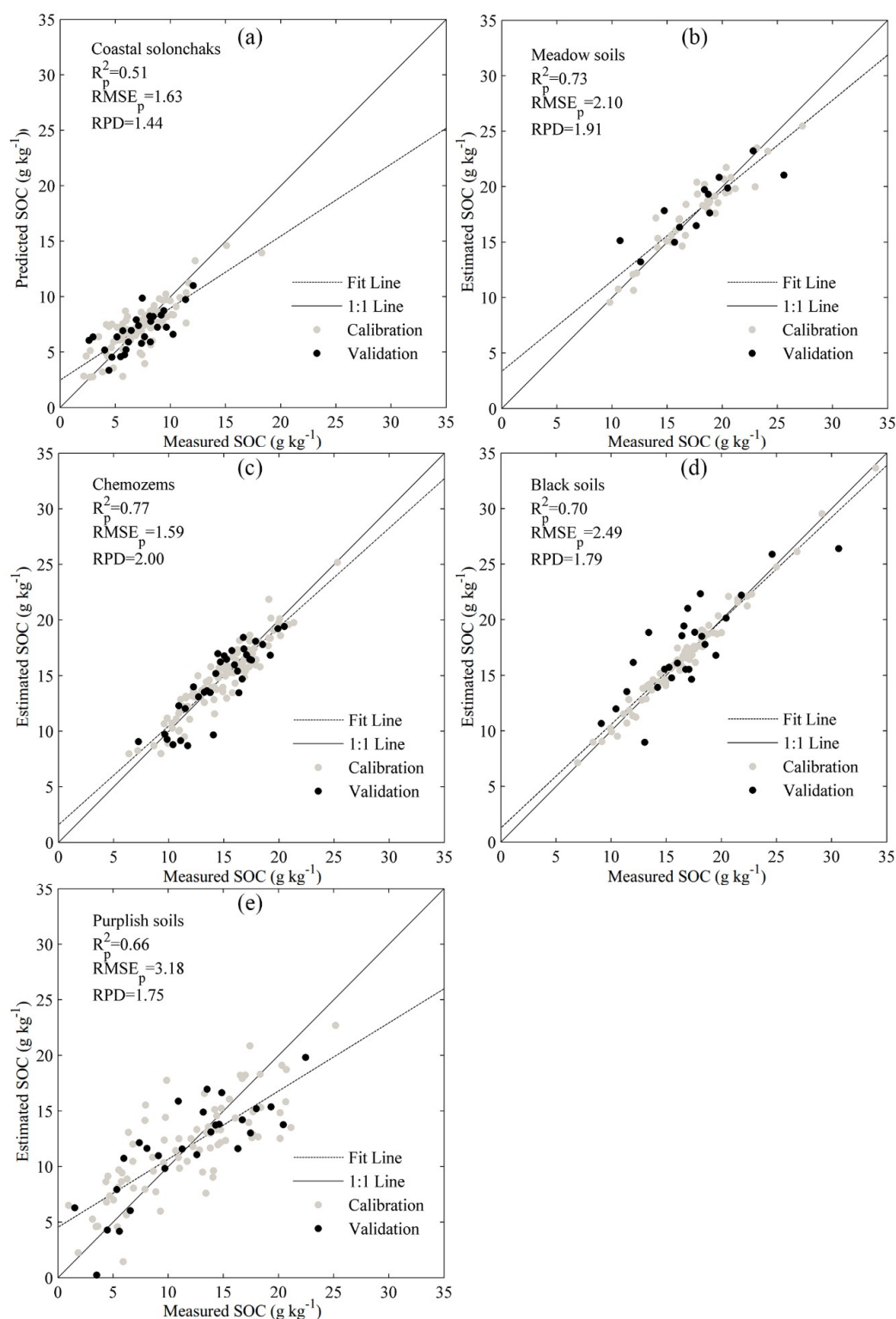


Figure 5. Estimated versus measured soil organic carbon (SOC) plots of spectroscopy models derived after discriminating five soil types in advance (Strategy II): Coastal solonchaks (a); Meadow soils (b); Chernozems (c); Black soils (d); and Purplish soils (e). R_p^2 denotes the coefficient of determination in prediction, RMSEP refers to the root mean square error of prediction, and RPD stands for residual predictive deviation.

4. Discussion

4.1. Soil Type Prediction through Vis-NIR Spectroscopy

Soil types can be accurately predicted with vis-NIR spectra and PLS-DA because of the absorption of vis-NIR spectroscopy through mineral and organic components. For example, coastal solonchaks are rich in salt and ions (Cl^- , Na^+ , and Ca^{2+}), purplish soils contain high level of CaCO_3 [64,65], and the three other soil types are high in SOC. Viscarra Rossel et al. [40] reviewed the important wavelengths used to predict soil types. In the current study, 86.72% of the validation samples were correctly predicted, which is similar to results in previous works [39,40,66].

Soil type was predicted using the spectra because information on the former is not always available in practical applications. Additional expert knowledge and cost are required to access soil types for a successful application of the soil library. Obtaining soil type by soil spectra overcomes this drawback, and shows reliable classification precision.

4.2. Effects of Stratifying Samples by Soil Type in SOC Estimation (Strategies II and III)

Stratifying samples in the soil library by soil type can improve the quality of SOC estimation models. In this study, the overall R_p^2 increased from 0.74 to 0.82–0.83 after stratification by soil type. Vasques et al. [14] obtained similar results; they stratified 6982 samples from Florida, USA, into seven soil orders, and found that the SOC models of all of the soil orders except Histosols are reliable. However, other researchers yielded different results. McDowell et al. [32] divided 307 samples of 10 soil types in the Hawaiian Islands into four broad soil groups and revealed that three soil groups did not exhibit an advantage over all of the samples. Madari et al. [33] separated 539 samples from Brazil into two soil orders and observed no improvement in the models.

Different conclusions were drawn because of the inappropriate comparisons between stratification and non-stratification techniques. The results of each soil type after stratification were compared with those of the entire dataset, disregarding that different validation sets were compared. For example, in the study of McDowell, the validation set of Andisol soils had 25–32 samples, whereas the entire dataset had 92 validation samples. To ensure that the validation set was comparable, we calculated the R_p^2 of the validation samples in each soil type, and the overall R_p^2 of all of the validation samples (Table 2). Our results suggested that stratification by soil type could improve the models for SOC estimation.

Stratification positively affected the SOC models because it produces homogeneous groups. Similar to previous findings [14], Welch's ANOVA showed SOC changes in relation to soil types ($p < 0.05$), indicating that the variance in SOC might be partly attributed to soil type. Applying a generic prediction model of SOM using all soil types is not desirable [38]. Another reason is the distinct spectral characteristics among soil types. Stratification by soil type also results in homogeneous groups in terms of spectral information. Thus, the homogeneity after stratification by soil type covered both spectra and SOC content. Shi et al. [41] divided the CSSL into five groups based on spectral clusters, and observed that the R^2 increased from 0.65 to 0.90. They [10] also considered geographical zones and spectral similarity, and observed homogeneous clusters. Other researchers performed clustering by using other variables, such as soil humidity, slope, parent material, and unsupervised Ward's Euclidian distance [25,28,31]. Clustering aims to correctly allocate the validation samples to the most similar group, and the SOC model based on that group can estimate the SOC of validation samples as accurately as possible. In most cases, the model that estimates soil properties is improved by clustering the samples into homogeneous groups [10,25,41].

4.3. Effects of Spectrally Derived Soil Type on SOC Estimation

Stratification by using spectrally derived soil type improves SOC estimation in a manner similar to that by using actual soil type. Mouazen et al. [67] speculated that the classification of samples by soil spectra into different texture classes can be used to establish separate models for each texture groups,

thereby improving the accuracy of vis-NIR spectroscopy models. To some extent, our results confirmed this assumption. Previous studies utilized soil type in soil property estimation through vis-NIR spectroscopy [14,32,33] or discriminated soil type through vis-NIR spectroscopy [39,40]. However, few reports have combined both procedures. The present study successfully proposed a strategy of including spectrally derived soil types into SOC estimation, and our results were satisfactory. This strategy required only the spectra and not the actual soil type of the validation samples.

While using the spectrally derived soil type, we encountered a problem regarding how misclassified samples affect the SOC models. A sample was misclassified because its spectral characteristics were more similar to those of the target soil type than to those of its actual soil type. In other words, the sample was allocated to a homogeneous group rather than its actual group in terms of spectral characteristics. For example, 38.46% of the samples in meadow soils were wrongly allocated, but the SOC model accuracy was slightly changed. For black soils, the SOC model that misclassified these samples was more suitable than their actual soil type. Misclassification slightly affected the vis-NIR estimation of SOC when stratified calibration strategies were applied.

The variable important projection (VIP) scores of the SOC models that were built for three soil types are shown in Figure 6 (VIP score analysis for the entire dataset is shown in Supplementary Materials Figure S2) to further investigate why misclassification slightly affects SOC estimation. The three soil types presented two different cases: no misclassification and severe misclassification. For coastal solonchaks, no sample was misclassified from or to the other soil types. By comparison, 30.44% of meadow soils were misclassified to chernozems, and 17.65% of chernozems were wrongly classified to meadow soils. The VIP score curve of meadow soils was similar to that of chernozems, indicating that the SOC estimation models of these two soil types exhibited some similarities. Thus, the misclassification of their samples would not result in differences in SOC prediction. The VIP score curves of coastal solonchaks were different from those of the two other soil types. If the samples of coastal solonchaks were misclassified to the two other soil types, then their influence on the subsequent SOC estimation was significant (Supplementary Materials Figure S2 and Table S2). Thus, the VIP score analysis confirmed that the homogeneous groups were similar to the SOC estimation models. Therefore, the influence of soil type misclassification through vis-NIR spectroscopy on SOC estimation was negligible.

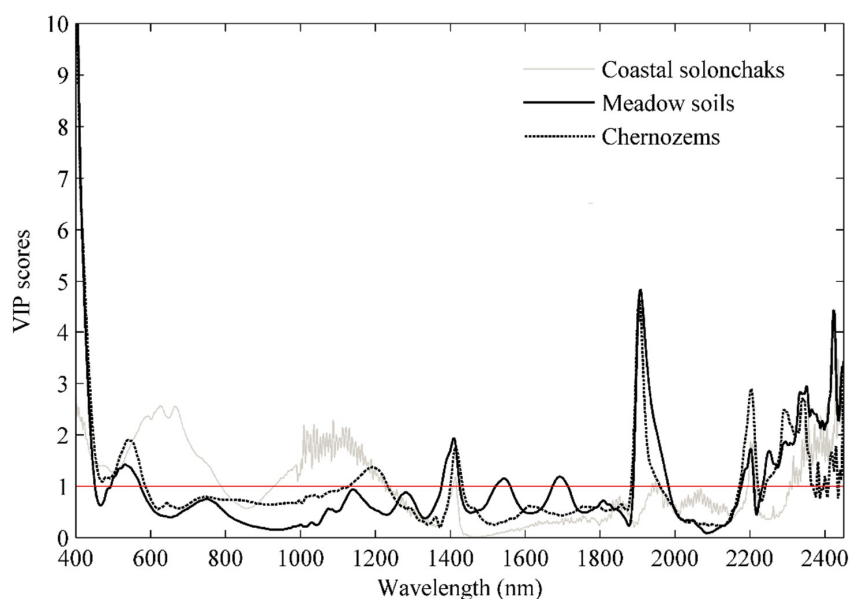


Figure 6. Variable importance projection (VIP) scores associated with the cross-validation of partial least squares regression model for soil organic carbon (SOC) concentration estimation through laboratory spectroscopy when samples were stratified by soil type. The threshold of VIP was set to 1 (red line).

5. Conclusions

Our study proposed a strategy of using the spectrally derived soil type as ancillary data to improve SOC estimation by utilizing vis-NIR spectroscopy and the Chinese soil library. The results allowed us to draw the following conclusions: (i) vis-NIR spectroscopy coupled with a soil library could be used for soil classification; (ii) stratifying samples by actual soil type (Strategy II) or spectrally derived soil type (Strategy III) significantly improved the quality of the SOC models for all of the soil types, and soil type was an adequate criterion for calibration set formation. The spectral misclassification of soil type in Strategy III slightly affected the robustness of the SOC estimation model, whereas Strategy III required low additional cost and was practically useful when soil classification was unavailable.

Despite our success of stratification by soil type, SOC estimation can still be improved using vis-NIR spectroscopy. Future study will be focused on other ancillary data, such as soil texture, pH, moisture, and land-use types, which might also be feasibly identified through vis-NIR spectroscopy and then included in SOC models. Our study focused on the CSSL, but our strategy could also be used in other countries or in a continental or global scale.

Supplementary Materials: The following are available online at <http://www.mdpi.com/2072-4292/10/11/1747/s1>, Figure S1: Location of the soil library with 515 samples in China. The location of samples from Meadow soils and Chernozems is unavailable. And some samples from Coastal solonchaks, Purplish soils and Black soils are also available. Figure S2: Variable importance projection (VIP) scores (black line) associated with the cross-validation of partial least-squares regression model for soil organic carbon concentration estimation by using laboratory spectroscopy and the entire dataset from Chinese soil spectral library. The threshold for VIP was set to 1 (horizontal dashed line). Figure S3: Scatter diagram of scores on latent variable 2 (LV2) plotted against latent variable 1 (LV1) for validation samples in partial least squares discriminant analysis (PLS-DA) models. The six samples in red ellipse were selected to be misclassified to Meadow soils and Chernozems. The other three ellipses (Bule, green, and dark green) were the 90% confidence ellipse for each soil type. Figure S4: Performance of SOC models stratified by soil type when the number of soil type varies from 5 to 12. Table S1: Spectral pretreatment for PLS-DA and PLSR. Table S2: The performance for the estimation models of soil organic carbon when six samples from Coastal solonchaks are misclassified to Meadow soils and Chernozems.

Author Contributions: Y.L. (Yi Liu), Y.C. and Y.L. (Yaolin Liu) conceived and designed the experiments. Y.L. (Yi Liu) and Y.C. analyzed the data. Z.S., G.Z., T.S., J.W. and Y.H. contributed greatly to data collection. Y.C. and S.L. reviewed and edited the draft. Y.L. (Yi Liu) wrote the paper. All authors read the submitted manuscript, and agreed to be listed as authors, and approved the version of manuscript for submission.

Funding: The APC was funded by the National Natural Science Foundation of China (Grant No. 41771440).

Acknowledgments: This study was supported by the National Natural Science Foundation of China (Grant No. 41771440, No. 41501444, and No. 41771432). We greatly appreciated the editors and the reviewers for their constructive suggestions and insightful comments which helped us greatly to improve this manuscript.

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

References

1. Lal, R. Soil carbon sequestration impacts on global climate change and food security. *Science* **2004**, *304*, 1623–1627. [[CrossRef](#)] [[PubMed](#)]
2. Lehmann, J.; Kleber, M. The contentious nature of soil organic matter. *Nature* **2015**, *528*, 60–68. [[CrossRef](#)] [[PubMed](#)]
3. Li, S.; Shi, Z.; Chen, S.; Ji, W.; Zhou, L.; Yu, W.; Webster, R. In situ measurements of organic carbon in soil profiles using vis-nir spectroscopy on the qinghai–tibet plateau. *Environ. Sci. Technol.* **2015**, *49*, 4980–4987. [[CrossRef](#)] [[PubMed](#)]
4. Nocita, M.; Stevens, A.; Toth, G.; Panagos, P.; van Wesemael, B.; Montanarella, L. Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. *Soil Biol. Biochem.* **2014**, *68*, 337–347. [[CrossRef](#)]
5. Wu, H.B.; Guo, Z.T.; Peng, C.H. Land use induced changes of organic carbon storage in soils of china. *Glob. Chang. Boil.* **2003**, *9*, 305–315. [[CrossRef](#)]
6. Bellon-Maurel, V.; McBratney, A. Near-infrared (nir) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils—critical review and research perspectives. *Soil Biol. Biochem.* **2011**, *43*, 1398–1410. [[CrossRef](#)]

7. Walkley, A.; Black, I.A. An examination of the degtjareff method for determining soil organic matter, and a proposed modification of the chromic acid titration method. *Soil Sci.* **1934**, *37*, 29–38. [[CrossRef](#)]
8. Nelson, D.; Sommers, L.E. Total carbon, organic carbon, and organic matter 1. *Methods Soil Anal. Part 2 Chem. Microbiol. Prop.* **1982**, 539–579.
9. Viscarra Rossel, R.A.; Hicks, W.S. Soil organic carbon and its fractions estimated by visible–near infrared transfer functions. *Eur. J. Soil Sci.* **2015**, *66*, 438–450. [[CrossRef](#)]
10. 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](#)]
11. Kuang, B.; Mouazen, A.M. Influence of the number of samples on prediction error of visible and near infrared spectroscopy of selected soil properties at the farm scale. *Eur. J. Soil Sci.* **2012**, *63*, 421–429. [[CrossRef](#)]
12. Viscarra Rossel, R.A.; Lobsey, C.R.; Sharman, C.; Flick, P.; McLachlan, G. Novel proximal sensing for monitoring soil organic c stocks and condition. *Environ. Sci. Technol.* **2017**, *51*, 5630–5641. [[CrossRef](#)] [[PubMed](#)]
13. Paz-Kagan, T.; Zaady, E.; Salbach, C.; Schmidt, A.; Lausch, A.; Zacharias, S.; Notesco, G.; Ben-Dor, E.; Karnieli, A. Mapping the spectral soil quality index (ssqi) using airborne imaging spectroscopy. *Remote. Sens.* **2015**, *7*, 15748–15781. [[CrossRef](#)]
14. Vasques, G.M.; Grunwald, S.; Harris, W.G. Spectroscopic models of soil organic carbon in florida, USA. *J. Environ. Qual.* **2010**, *39*, 923–934. [[CrossRef](#)] [[PubMed](#)]
15. Ji, W.; Li, S.; Chen, S.; Shi, Z.; Viscarra Rossel, R.A.; Mouazen, A.M. Prediction of soil attributes using the chinese soil spectral library and standardized spectra recorded at field conditions. *Soil Tillage Res.* **2016**, *155*, 492–500. [[CrossRef](#)]
16. Viscarra Rossel, R.A.; Behrens, T.; Ben-Dor, E.; Brown, D.; Demattê, J.; Shepherd, K.; Shi, Z.; Stenberg, B.; Stevens, A.; Adamchuk, V.; et al. A global spectral library to characterize the world’s soil. *Earth-Sci. Rev.* **2016**, *155*, 198–230. [[CrossRef](#)]
17. Gogé, F.; Gomez, C.; Jolivet, C.; Joffre, R. Which strategy is best to predict soil properties of a local site from a national vis–nir database? *Geoderma* **2014**, *213*, 1–9. [[CrossRef](#)]
18. 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](#)]
19. Brown, D.J.; Shepherd, K.D.; Walsh, M.G.; Mays, M.D.; Reinsch, T.G. Global soil characterization with vnir diffuse reflectance spectroscopy. *Geoderma* **2006**, *132*, 273–290. [[CrossRef](#)]
20. Viscarra Rossel, R.A.; Behrens, T. Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma* **2010**, *158*, 46–54. [[CrossRef](#)]
21. Ji, W.; Li, X.; Li, C.; Zhou, Y.; Shi, Z. Using different data mining algorithms to predict soil organic matter based on visible-near infrared spectroscopy. *Spectrosc. Spectr. Anal.* **2012**, *32*, 2393–2398.
22. Gholizadeh, A.; Borůvka, L.; Saberioon, M.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.* **2015**, *10*, 218–227. [[CrossRef](#)]
23. Rammal, A.; Perrin, E.; Vrabie, V.; Bertrand, I.; Habrant, A.; Chabbert, B. Optimal preprocessing and FCM clustering of MIR, NIR and combined MIR-NIR spectra for classification of maize roots. In Proceedings of the 2014 Third International Conference on e-Technologies and Networks for Development (ICeND), Beirut, Lebanon, 29 April–1 May 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 110–115.
24. Gras, J.-P.; Barthès, B.G.; Mahaut, B.; Trupin, S. Best practices for obtaining and processing field visible and near infrared (VNIR) spectra of topsoils. *Geoderma* **2014**, *214*, 126–134. [[CrossRef](#)]
25. 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](#)]
26. Igne, B.; Reeves, J.B.; McCarty, G.; Hively, W.D.; Lund, E.; Hurburgh, C.R. Evaluation of spectral pretreatments, partial least squares, least squares support vector machines and locally weighted regression for quantitative spectroscopic analysis of soils. *J. Near Infrared Spectrosc.* **2010**, *18*, 167–176. [[CrossRef](#)]
27. Gogé, F.; Joffre, R.; Jolivet, C.; Ross, I.; Ranjard, L. Optimization criteria in sample selection step of local regression for quantitative analysis of large soil NIRS database. *Chemom. Intell. Lab. Syst.* **2012**, *110*, 168–176. [[CrossRef](#)]

28. Cierniewski, J.; Kaźmierowski, C.; Kuśnierek, K.; Piekarczyk, J.; Kwlewicz, S.; Guliński, M.; Terelak, H.; Stuczyński, T.; Maliszewska-Kordybach, B. Unsupervised clustering of soil spectral curves to obtain their stronger correlation with soil properties. In Proceedings of the 2010 2nd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), Reykjavik, Iceland, 14–16 June 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 1–4.
29. Udelhoven, T.; Emmerling, C.; Jarmer, T. Quantitative analysis of soil chemical properties with diffuse reflectance spectrometry and partial least-square regression: A feasibility study. *Plant Soil* **2003**, *251*, 319–329. [[CrossRef](#)]
30. Stevens, A.; Udelhoven, T.; Denis, A.; Tychon, B.; Liroy, R.; Hoffmann, L.; Van Wesemael, B. Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy. *Geoderma* **2010**, *158*, 32–45. [[CrossRef](#)]
31. Peng, Y.; Knadel, M.; Gislum, R.; Deng, F.; Norgaard, T.; de Jonge, L.W.; Moldrup, P.; Greve, M.H. Predicting soil organic carbon at field scale using a national soil spectral library. *J. Near Infrared Spectrosc.* **2013**, *21*, 213–222. [[CrossRef](#)]
32. McDowell, M.L.; Bruland, G.L.; Deenik, J.L.; Grunwald, S. Effects of subsetting by carbon content, soil order, and spectral classification on prediction of soil total carbon with diffuse reflectance spectroscopy. *Appl. Environ. Soil Sci.* **2012**, *2012*, 294121. [[CrossRef](#)]
33. Madari, B.E.; Reeves, J.B., III; Coelho, M.R.; Machado, P.L.; De-Polli, H.; Coelho, R.M.; Benites, V.M.; Souza, L.F.; McCarty, G.W. Mid-and near-infrared spectroscopic determination of carbon in a diverse set of soils from the brazilian national soil collection. *Spectrosc. Lett.* **2005**, *38*, 721–740. [[CrossRef](#)]
34. Brown, D.J. Using a global VNIR soil-spectral library for local soil characterization and landscape modeling in a 2nd-order uganda watershed. *Geoderma* **2007**, *140*, 444–453. [[CrossRef](#)]
35. Guerrero, C.; Zornoza, R.; Gómez, I.; Mataix-Beneyto, J. Spiking of NIR regional models using samples from target sites: Effect of model size on prediction accuracy. *Geoderma* **2010**, *158*, 66–77. [[CrossRef](#)]
36. Guerrero, C.; Stenberg, B.; Wetterlind, J.; Viscarra Rossel, R.A.; Maestre, F.; Mouazen, A.M.; Zornoza, R.; Ruiz-Sinoga, J.; Kuang, B. Assessment of soil organic carbon at local scale with spiked NIR calibrations: Effects of selection and extra-weighting on the spiking subset. *Eur. J. Soil Sci.* **2014**, *65*, 248–263. [[CrossRef](#)]
37. Wetterlind, J.; Stenberg, B. Near-infrared spectroscopy for within-field soil characterization: Small local calibrations compared with national libraries spiked with local samples. *Eur. J. Soil Sci.* **2010**, *61*, 823–843. [[CrossRef](#)]
38. Ogen, Y.; Neumann, C.; Chabrilat, S.; Goldshleger, N.; Ben Dor, E. Evaluating the detection limit of organic matter using point and imaging spectroscopy. *Geoderma* **2018**, *321*, 100–109. [[CrossRef](#)]
39. Vasques, G.M.; Dematte, J.A.M.; Viscarra Rossel, R.A.; Ramirez-Lopez, L.; Terra, F.S. Soil classification using visible/near-infrared diffuse reflectance spectra from multiple depths. *Geoderma* **2014**, *223*, 73–78. [[CrossRef](#)]
40. Viscarra Rossel, R.A.; Webster, R. Discrimination of australian soil horizons and classes from their visible–near infrared spectra. *Eur. J. Soil Sci.* **2011**, *62*, 637–647. [[CrossRef](#)]
41. 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](#)]
42. Zhang, W.L.; Ai-Guo, X.U.; Zhang, R.L.; Hong, J.I. Review of soil classification and revision of china soil classification system. *Sci. Agric. Sin.* **2014**, *47*, 3214–3230.
43. IUSS Working Group Wrb. *World Reference Base for Soil Resources 2014 International Soil Classification System for Naming Soils and Creating Legends for Soil Maps*; FAO: Rome, Italy, 2014.
44. 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](#)]
45. Zhang, J.; Xu, Z. Dye tracer infiltration technique to investigate macropore flow paths in maka mountain, yunnan province, china. *J. Central South Univ.* **2016**, *23*, 2101–2109. [[CrossRef](#)]
46. Rozenstein, O.; Paz-Kagan, T.; Salbach, C.; Karnieli, A. Comparing the effect of preprocessing transformations on methods of land-use classification derived from spectral soil measurements. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 2393–2404. [[CrossRef](#)]
47. Debaene, G.; Niedzwiecki, J.; Pecio, A.; Zurek, A. Effect of the number of calibration samples on the prediction of several soil properties at the farm-scale. *Geoderma* **2014**, *214*, 114–125. [[CrossRef](#)]

48. Steinier, J.; Termonia, Y.; Deltour, J. Smoothing and differentiation of data by simplified least square procedure. *Anal. Chem.* **1972**, *44*, 1906–1909. [[CrossRef](#)] [[PubMed](#)]
49. Stenberg, B.; Viscarra Rossel, R.A.; Mouazen, A.M.; Wetterlind, J. Chapter five-visible and near infrared spectroscopy in soil science. *Adv. Agron.* **2010**, *107*, 163–215.
50. Naes, T.; Martens, H. *Multivariate Calibration*; Norwegian Food Research Institute: Ås, Norway, 1989.
51. Barnes, R.; Dhanoa, M.; Lister, S.J. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl. Spectrosc.* **1989**, *43*, 772–777. [[CrossRef](#)]
52. Moros, J.; Fdez-Ortiz, D.V.S.; Gredilla, A.; De, D.A.; Madariaga, J.M.; Garrigues, S.; De, L.G.M. Use of reflectance infrared spectroscopy for monitoring the metal content of the estuarine sediments of the nerbioi-ibaizabal river (metropolitan bilbao, bay of biscay, basque country). *Environ. Sci. Technol.* **2009**, *43*, 9314–9320. [[CrossRef](#)] [[PubMed](#)]
53. Viscarra Rossel, R.A. Parles: Software for chemometric analysis of spectroscopic data. *Chemom. Intell. Lab. Syst.* **2008**, *90*, 72–83. [[CrossRef](#)]
54. Echambadi, R.; Hess, J.D. Mean-centering does not alleviate collinearity problems in moderated multiple regression models. *Mark. Sci.* **2007**, *26*, 438–445. [[CrossRef](#)]
55. Kuhn, M.; Johnson, K. *Data pre-processing. Applied Predictive Modeling*; Springer: Berlin, Germany, 2013; pp. 27–59.
56. Li, Z.; Liu, J.; Shan, P.; Peng, S.; Lv, J.; Ma, Z. Strategy for constructing calibration sets based on a derivative spectra information space consensus. *Chemom. Intell. Lab. Syst.* **2016**, *156*, 7–13. [[CrossRef](#)]
57. Goodarzi, M.; Sharma, S.; Ramon, H.; Saeys, W. Multivariate calibration of NIR spectroscopic sensors for continuous glucose monitoring. *TrAC Trends Anal. Chem.* **2015**, *67*, 147–158. [[CrossRef](#)]
58. Della Riccia, G.; Del Zotto, S. A multivariate regression model for detection of fumonisins content in maize from near infrared spectra. *Food Chem.* **2013**, *141*, 4289–4294.
59. Hazama, K.; Kano, M. Covariance-based locally weighted partial least squares for high-performance adaptive modeling. *Chemom. Intell. Lab. Syst.* **2015**, *146*, 55–62. [[CrossRef](#)]
60. Minasny, B.; McBratney, A. Why you don't need to use RPD. *Pedometron* **2013**, *33*, 14–15.
61. Bellon-Maurel, V.; Fernandez-Ahumada, E.; Palagos, B.; Roger, J.-M.; McBratney, A. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. *TrAC Trends Anal. Chem.* **2010**, *29*, 1073–1081. [[CrossRef](#)]
62. Hong, Y.; Yu, L.; Chen, Y.; Liu, Y.; Liu, Y.; Liu, Y.; Cheng, H. Prediction of soil organic matter by VIS–NIR spectroscopy using normalized soil moisture index as a proxy of soil moisture. *Remote Sens.* **2017**, *10*, 28. [[CrossRef](#)]
63. Szymańska, E.; Saccenti, E.; Smilde, A.K.; Westerhuis, J.A. Double-check: Validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics* **2012**, *8*, 3–16. [[CrossRef](#)] [[PubMed](#)]
64. Li, L.; Xia, J.; Liu, L. Effect of purple soil organic matter on adsorption and desorption of Pb²⁺ by aggregates. *Chin. J. Ecol.* **2014**, *33*, 1274–1283.
65. Liu, Y.; Pan, X.; Wang, C.; Li, Y.; Zhou, R.; Xie, X.; Wang, M. Prediction of coastal saline soil salinity based on VIS-NIR reflectance spectroscopy. *Acta Pedofil Sin.* **2012**, *49*, 824–829.
66. Du, C.; Linker, R.; Shaviv, A. Identification of agricultural mediterranean soils using mid-infrared photoacoustic spectroscopy. *Geoderma* **2008**, *143*, 85–90. [[CrossRef](#)]
67. Mouazen, A.M.; Karoui, R.; De Baerdemaeker, J.; Ramon, H. Classification of soil texture classes by using soil visual near infrared spectroscopy and factorial discriminant analysis techniques. *J. Near Infrared Spectrosc.* **2005**, *13*, 231–240. [[CrossRef](#)]

