forests

MDPI

Article

Improved Soil Organic Carbon Prediction in a Forest Area by Near-Infrared Spectroscopy: Spiking of a Soil Spectral Library

Miao Long [1,2], Tianxiang Yue [1,3,4,5,6], Zhe Xu [3], Jiaxin Guo [1,2], Jie Luo [6], Xi Guo [1,2] and Xiaomin Zhao [1,2,*]

1 College of Land Resources and Environment, Jiangxi Agricultural University, Nangchang 330045, China
2 Key Laboratory of Poyang Lake Watershed Agricultural Resources and Ecology of Jiangxi Province, Nanchang 330045, China
3 State Key Laboratory of Resources and Environment Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
4 College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100190, China
5 Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China
6 School of Resource, Environment and Jewelry, Jiangxi College of Applied Technology, Ganzhou 341000, China
* Correspondence: zhaoxm889@126.com

Abstract: The rapid quantitative assessment of soil organic carbon (SOC) is essential for understanding SOC dynamics and developing management strategies in forest ecosystems. Compared with traditional laboratory methods, visible and near-infrared spectroscopy is an efficient and inexpensive technique widely used to predict SOC content. Herein, we compared three different spiking strategies. That is, a large-scale global soil spectral library (global-SSL; 3122 samples) was used as the basis for predicting SOC content in a small-scale local soil spectral library (local-SSL; 89 samples) in Wugong Mountain, Jiangxi Province, China. Partial least squares regression models using global-SSL 'spiking' with local samples did not necessarily achieve more accurate predictions than models using local-SSL. Using the developed strategy, a calibration set can be established by selecting the top *N* spectral samples from global-SSL with high similarity to each local sample, together with the 'spiking' set from local-SSL. It is possible to individually improve the prediction results based on local samples (R^2 = 0.90, RMSE = 7.19, RPD = 3.38) and still allow for quantitative prediction from fewer local calibration samples (R^2 = 0.83, RMSE = 8.71, RPD = 2.68). The developed method is cost-effective and accurate for local-scale SOC assessment in target forest areas using a large soil spectral library.

Keywords: near-infrared spectroscopy; soil organic carbon; soil spectral library; spiking; forest assessment

Citation: Long, M.; Yue, T.; Xu, Z.; Guo, J.; Luo, J.; Guo, X.; Zhao, X. Improved Soil Organic Carbon Prediction in a Forest Area by Near-Infrared Spectroscopy: Spiking of a Soil Spectral Library. *Forests* **2023**, *14*, 118. https://doi.org/10.3390/f14010118

Academic Editor: Arturo Sanchez-Azofeifa

Received: 28 November 2022
Revised: 4 January 2023
Accepted: 5 January 2023
Published: 8 January 2023

1. Introduction

Forest ecosystems, which account for 31% of the global land area, are the main reserve of terrestrial carbon (C) stock. Given the large area and wide distribution of forests, forest soil plays an essential role in the global C cycle [1]. Monitoring soil organic C (SOC) dynamics in forests is necessary to address key environmental issues, such as improving soil health and mitigating climate change [2]. However, traditional SOC measurement methods are expensive and time-consuming, with complex analytical procedures [3]. The development of reliable, accurate, rapid, and cost-effective methods for quantitative SOC assessment would greatly assist in forest soil management.

Based on the relationship between soil spectral reflectance and the spectral response of soil organic matter, visible and near-infrared (vis–NIR) spectroscopy has proven to be a rapid, accurate, and cost-effective method for predicting SOC content [4]. Numerous studies have reported on local-scale SOC prediction by vis-NIR spectroscopy [5–8], and these predictions are usually accurate. The samples in the calibration and validation sets have similar soil properties (e.g., parent material and soil type). However, it is not always

feasible to measure soil properties in new target areas using existing locally calibrated models because the prediction accuracy may be substantially reduced, and the cost of modeling each study area separately is high [9–12].

Over the last decade, great efforts have been dedicated to building a large soil spectral library covering as many soil types and properties as possible and developing modeling strategies for new target areas [13]. Soil spectral libraries are developed at different scales, ranging from field to local, national, continental, and global, while some are freely and publicly available for research use [10,14–19]. However, predictions are often inaccurate or biased when local calibration models are constructed using large soil spectral libraries because the spectral characteristics of the local soil may not be appropriately reflected in the calibration [10,20]. The most common strategies to improve the accuracy of regional or local calibrations using soil spectral libraries are (i) selecting optimal calibration samples from the soil spectral library and (ii) spiking the soil spectral library with representative local samples.

To select optimal soil samples as the calibration set, researchers generally search the soil spectral library for samples with similar properties (e.g., spectra, soil type, and land use pattern) compared to local soils [4,14,18,21–24]. When spiking the soil spectral library, several representative local samples are added to the calibration set to correctly reflect the local soil properties. The main objective of spiking is to improve the prediction of soil properties by including a small number of local samples in the calibration set [9,18,25–28]. The number of calibration samples also affects the prediction performance [29]; the larger the number of local samples included in the 'spiking' set, the higher the prediction accuracy of the soil properties [26,28]. However, setting the size of the 'spiking' set to large can reduce the benefits of vis-NIR spectroscopy as an affordable and rapid analytical method. On the other hand, setting the size of 'spiking' too small can reduce the amount of important information available for modeling, resulting in less stable calibration [9].

A large soil spectral library usually consists of samples associated with different soil-forming factors (e.g., geographic settings, parent materials, and vegetation types), which are considered heterogeneous. By contrast, soil samples from a target area are formed with similar soil-forming factors. The local samples are a small regional homogeneous collection, as revealed by vis-NIR spectroscopy and physicochemical analyses [30]. Although it is advantageous to estimate soil properties after the stratification of soil spectral libraries using soil types [4], field sampling imposes an additional cost burden and requires expert knowledge to make accurate judgments about soil types. Similar samples can be selected from a large soil spectral library and a local one based on the soil spectra only [31]. Several studies have analyzed soil spectral similarities by using the spectral angle mapper (SAM) algorithm [32–35]. SAM compares spectral similarities by calculating the angle between samples in spectral space [36]. This algorithm is easy to execute and insensitive to the effects of light and spectral reflectance magnitude, which facilitates in situ soil monitoring [37].

In the present study, we verified the feasibility of accurately predicting the local SOC content in a target forest area using a global soil spectral library (global-SSL) spiked with samples from a local soil spectral library (local-SSL). The objective of the study was to compare various spiking strategies and determine the optimal strategy for local SOC prediction in forests using global-SSL. We compared three modeling strategies: (i) a modeling approach without a 'spiking' set, involving different numbers of local samples in the calibration sets; (ii) a modeling approach involving a random selection of samples from global-SSL as a calibration subset, combined with the 'spiking' set from local-SSL; and (iii) a modeling approach involving the strategic selection of samples from global-SSL as a calibration subset, along with the 'spiking' set from local-SSL.

2. Materials and Methods

2.1. Soil Sampling and Laboratory Measurements

This study was conducted in Wugong Mountain in the northwestern Jiangxi Province, China (27°24′–27°34′ N, 114°05′–114°15′ E). The study area has steep, vertical mountains with high altitudes, with the highest point at Baihe Peak (Jin Ding) reaching 1918.3 m above sea level. The vertical zonation of climate, soil, and vegetation is distinct here. The average annual temperature is 14–16 °C, the highest summer temperature is 23 °C, the average annual sunshine duration is 1580–1700 h, the average annual evaporation is 1360–1700 mm, the average annual humidity value is 70%–80%, and the average annual precipitation is 1350–1570 mm. The vegetation of Wugong Mountain shows typical vertical zonation, and the whole vertical zonation of the mountain shows the type of evergreen broad-leaved forest that exists in the middle subtropics, which mainly shows an evergreen broad-leaved forest, mixed evergreen deciduous broad-leaved forest, coniferous forest, mountain elfin forest, bamboo forest, shrubbery, and meadow from the foot to the top of the mountain. The rock types in the area are mainly granite, gneiss, phyllite, and arenite. The soil type at this site is classified as Cambisols in the World Reference Base for Soil Resources (WRB). There are abundant dead branches and leaves, yet the decomposition of organic matter is slow due to local temperature and moisture conditions. Consequently, the soil is characterized by a dark color and high organic matter content.

We collected soil samples from the study area in September 2021. Representative sampling locations were selected by observing the surrounding environmental conditions (e.g., slopes, ridges, backs of the mountains, and gullies) within an area with an elevation difference of ~50 m and a certain horizontal distance. Five topsoil samples (depth 0–20 cm) were randomly collected from each sampling location and mixed to form a composite sample, which could reduce the sampling error and better reflect the average level of the sampling location. A total of 89 soil samples were collected, and each sample was placed in a re-sealable bag to facilitate storage and prevent cross-contamination. The longitude, latitude, and altitude of each sampling location were obtained by using a GPSMAP 669S handheld global positioning system (Garmin, Taiwan, China), while its environmental characteristics were recorded.

Fresh soil samples were transported to the laboratory, where they were air-dried. After manually removing identifiable non-soil impurities, the samples were mixed thoroughly and reduced to 500 g by quartering. The samples were ground and divided into two portions; one portion was sieved through a 2 mm sieve for spectral measurements, and the other was sieved through a 0.149 mm sieve for SOC determination. SOC content was analyzed using the potassium dichromate volumetric method with external heating according to the Forestry Industry Standard of China—Determination of Soil Available Nutrients (LY/T 1237—2015). Under external heating conditions (oil bath at 180 °C, boiling for 5 min), the SOC was oxidized using a standard solution of potassium dichromate as an oxidant, and the excess potassium dichromate was titrated with a standard solution of divalent iron, and the SOC content could be calculated from the amount of potassium dichromate consumed. Compared to the dry burning method, the results measured by this method could only oxidize 90% of the SOC, so the final oxidation correction factor of 1.1 was multiplied to calculate the SOC content.

2.2. Spectra Acquisition and Data Pre-Processing

Soil spectral reflectance was measured using an ASD FieldSpec4 Pro FR spectrometer (ASD Inc., Boulder, CO, USA) in the wavelength range of 350–2500 nm. The instrument was warmed up for 30 min before data acquisition. It was calibrated with a 99% reflectance Spectralon panel (Labsphere, North Sutton, NH, USA) before each scan to reduce measurement errors. Spectral measurements were repeated 10 times for each sample and averaged to generate a dataset [38].

Pre-processing spectra can reduce the influence of environmental factors and interference from the instrument system noise, thus yielding spectral data with a high signal-

to-noise ratio and improving the stability and accuracy of the model. We corrected the spectral data for splice points with abrupt spectral changes between different detectors using the Splice Correction function in ViewSpec Pro software (version 6.0; Analytical Spectral Devices, Malvern Panalytical, Boulder, CO, USA). In addition, the spectrum between 400 nm and 2450 nm was retained, and edge bands (350–399 nm, 2451–2500 nm) with low signal-to-noise ratios were excluded to reduce the interference from high-frequency noise [39]. Each reflectance spectrum was resampled by selecting every 10^{th} nanometer to reduce the dimensionality so that each spectrum consisted of 205 bands per sample.

2.3. Description of the Two Soil Spectral Libraries

The global-SSL used in this study is based on a soil spectral library developed by the World Agroforestry Center (ICRAF; http://www.isric.org/data/ (accessed on 14 June 2022)). The spectral library consists of 4437 samples from 58 countries spanning Africa, Asia, Europe, North America, and South America. Among them, 3122 samples containing both SOC and spectral data were selected as global-SSL. The SOC content was analyzed using the Walkley–Black method (World Agroforestry (ICRAF) and International Soil Reference Information Centre (ISRIC)—World Soil Information System, 2010). SOC and spectral data for 89 samples collected from Wugong Mountain were used as local-SSL.

2.4. Partial Least Squares Regression

Partial least squares regression (PLSR) is a powerful modeling tool that integrates the features of principal component analysis and multiple regression. It reduces a large number of measured cross-tabulated spectral variables to a small number of uncorrelated latent variables [40]. PLSR is widely used in chemometrics and the quantitative analysis of soil spectra, especially in the local range [41–43]. It also has a wide application in developing local and regional calibration models for SOC prediction from soil spectra [25]. In the present study, the optimal number of latent variables used for the PLSR model was selected according to the lowest root mean square error (RMSE) in 10-fold cross-validation [44].

2.5. Modeling Methods and Strategies

To determine the best calibration model for SOC prediction, three different strategies were employed to build a useful and cost-effective model for predicting SOC content in local-SSL (Figure 1). First, we used the Kennard–Stone algorithm to divide the 89 local samples into a calibration set and a validation set [45]. Rather than a random selection, this division was chosen to ensure that both the calibration and validation sets would cover the spectral diversity in local-SSL. The calibration set consisted of 63 samples, and the validation set contained 26 samples, with a 7:3 ratio based on Euclidean distances from each other (Set1). We also set the calibration and validation sets ratio in local-SSL to 53/36 (Set2) and 43/46 (Set3). This was conducted to verify the possibility of predicting more unknown samples by using fewer local-SSL samples as a 'spiking' subset and selecting global-SSL samples as the calibration set. To facilitate a comparison of the prediction performance of the different modeling strategies, the same validation sets (Set1/Set2/Set3) were used for all three strategies; the 'spiking' set used in Strategies II and III was the same as the calibration sets (Set1/Set2/Set3) used in Strategy I.

Next, we used SAM to select samples from global-SSL that were spectrally similar to samples in local-SSL as the calibration subset. Among the three proposed modeling strategies, Strategy I uses samples from local-SSL to build the model, hereafter referred to as the 'no-spiking strategy'. Strategies II and III use samples from local-SSL as the 'spiking' subset and samples from global-SSL as the calibration subset, hereafter referred to as the 'spiking strategy'.

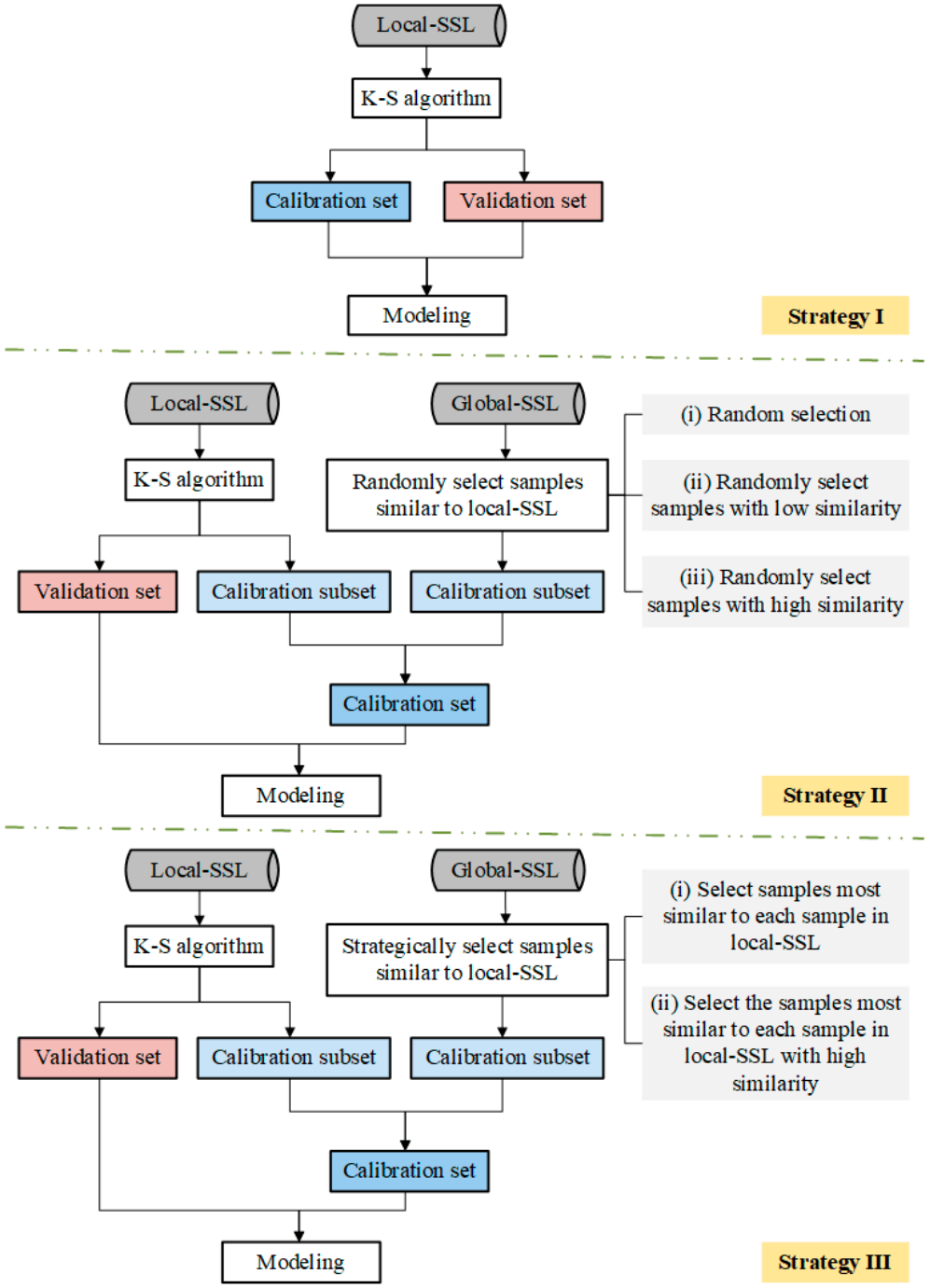

Figure 1. Flowchart explaining the information in and differences between spectral datasets (calibration/validation) and modeling strategies used in this study.

- Strategy I: Modeling approach with the 'no-spiking strategy'

We used only local-SSL in the calibration set selection step to establish Calibration Set1, Set2, and Set3. This was conducted to test the predictive performance of the model using selected samples from local-SSL as a calibration set without a 'spiking' set.

- Strategy II: Modeling approach with the 'spiking strategy' (random selection of global-SSL samples as a calibration subset)

We randomly selected the calibration subset using both local-SSL and global-SSL in the calibration set selection step. To compare the advantages and disadvantages of

the random samples selected from global-SSL as calibration subsets, the following three different selection methods were used: (i) the unconstrained random selection of N samples from global-SSL as calibration subsets; (ii) the random selection of N samples from global-SSL with low spectral similarity to each sample in the local-SSL as calibration subsets; and (iii) the random selection of N samples from global-SSL with high spectral similarity to some samples in the local-SSL as calibration subsets. The calibration set of the model consisted of the calibration subset from global-SSL and the 'spiking' set from local-SSL (Calibration Set1/Set2/Set3). To reduce the influence of random factors, modeling was repeated 100 times based on the random selection of samples, and the results were averaged.

- Strategy III: Modeling approach with the 'spiking strategy' (strategic selection of global-SSL samples as a calibration subset)

We strategically selected the calibration subset using both local-SSL and global-SSL in the calibration set selection step. Two selection methods were used: (i) selecting the top N spectral samples from global-SSL that were the most similar to each sample in local-SSL as the calibration subset, and (ii) selecting the high-similarity samples based on method (i) as the calibration subset. The calibration set of the model consisted of the calibration subset from global-SSL and the 'spiking' set from local-SSL (Calibration Set1/Set2/Set3). Here, N was artificially selectable, and the number of calibration samples extracted from global-SSL was uncertain. The N value was based on the spectral similarity of the samples in global-SSL and local-SSL and was not artificially controllable.

For example, there were m samples in the local-SSL, and if the top N samples from global-SSL with the highest spectral similarity to each local sample were selected as the calibration subset, theoretically, there would be $m \times N$ samples selected into the calibration subset. Owing to the similarity between the samples in local-SSL, a sample in global-SSL was inevitably similar to multiple samples in local-SSL, and the actual number of samples selected into the calibration subset would be less than or equal to $m \times N$. Despite its uncertainty, this sample number ensured that the most similar samples of each sample in local-SSL were selected from global-SSL as a calibration subset.

Several combinations of calibration and validation sets for the different modeling strategies (Figure 1) were used to construct PLSR models. All the pre-processing and modeling steps were performed using MATLAB (version R2021a; The MathWorks Inc., Natick, MA, USA).

2.6. Model Evaluation

Three statistical metrics were used to evaluate the SOC prediction models developed with different spiking strategies. The coefficient of determination (R^2) reflects the stability of the model: the closer R^2 is to 1, the better the stability and goodness-of-fit the model demonstrates [46]. RMSE indicates the deviation between the predicted value and the original data; the lower the RMSE value, the smaller the prediction error and the higher the accuracy [47]. The calibration performance of the two methods was compared by using the ratio of performance to deviation (RPD), which is the ratio of the standard deviation of the original data to the RMSE of validation. RPD values were divided to indicate six levels of prediction accuracy as follows: RPD < 1.0 indicates very poor predictions, and their use is not recommended; 1.0 < RPD < 1.4 indicates poor predictions where only high and low values are distinguishable; 1.4 < RPD < 1.8 indicates fair predictions which may be used for assessment and correlation; 1.8 < RPD < 2.0 indicates good predictions where quantitative predictions are possible; 2.0 < RPD < 2.5 indicates very good, quantitative predictions; and RPD > 2.5 indicates excellent predictions [48]. In contrast to RPD, the ratio of performance to InterQuartile distance (RPIQ), which used interquartile ranges rather than standard deviations to account for the population distribution of the skewed data sets, is also widely used as an evaluation metric, with larger values indicating better model prediction.

The three statistical metrics were computed according to the following Equations (1)–(4):

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_{mi} - y_{pi})^2}{\sum_{i=1}^{n} (y_{mi} - \overline{y}_m)^2} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_{mi} - y_{pi})^2} \quad (2)$$

$$RPD = SD/RMSE_V \quad (3)$$

$$RPIQ = (Q_3 - Q_1)/RMSE_v \quad (4)$$

where n is the number of samples, y_{mi} is the measured SOC content of sample i, y_{pi} is the predicted SOC content of sample i, $\overline{y}_m$ is the mean of the measured SOC content, SD is the standard deviation of the measured SOC content, and $RMSE_v$ is the RMSE of the validation set. Q_3 is the third quartile of the SOC content, and Q_1 is the first quartile of the SOC content. Descriptive statistical analysis was performed using SPSS Statistics (version 20; IBM Corp., Armonk, NY, USA).

3. Results and Discussion

3.1. Descriptive Statistics of the SOC Content

The descriptive statistics of the measured SOC content in the study area are summarized in Table 1. In local-SSL, SOC levels ranged from 18.44 to 134.32 g·kg^{-1}, with a mean value of 56.12 g·kg^{-1} and a standard deviation of 25.54 g·kg^{-1}. There was an excess kurtosis value of 1.10, which indicates a slight tendency for outliers in the dataset. The skewness was 1.23, which demonstrates that the data were skewed and normally distributed. Due to geospatial proximity, soil samples in local-SSL were developed under similar soil-forming conditions (e.g., parent material and vegetation type), with a minor variation in soil properties [49].

Table 1. Statistics for soil organic carbon content (g·kg^{-1}) measured in the laboratory.

Soil Spectral Library	*N*	Min	Median	Mean	Max	SD	Ske	EK
Local-SSL	89	18.44	48.94	56.12	134.32	25.54	1.23	1.10
Global-SSL	3122	0.10	0.30	13.69	627.80	13.68	10.23	137.66

Min—Minimum; Max—Maximum; SD—standard deviation; Sk—Skewness; EK—Excess kurtosis.

In global-SSL, a large variation in SOC content was observed among the samples (0.10–627.80 g·kg^{-1}), with a mean value of 13.69 g·kg^{-1} and a standard deviation of 13.68 g·kg^{-1}. The skewness and excess kurtosis of the SOC levels were as high as 10.23 and 137.66, respectively. This is due to the fact that global-SSL contains soil spectra from different locations in multiple countries. In both spectral libraries, the distribution of the SOC content was concentrated below the mean value.

3.2. Soil Spectra

Figure 2 shows that the SOC content was divided into six groups according to the level of <35 (a), 35–50 (b), 50–65 (c), 65–80 (d), 80–95 (c), and >95 (f) g·kg^{-1}. The local dataset had a high SOC content. Stoner et al. classified soil spectral curves into five types, and this soil spectral curve type were the organic-dominated form [50]. The soil spectral reflectance decreased as the SOC content increased. The shapes of the spectral curves for different SOC contents did not differ prominently with the increasing SOC content. However, the SOC content <35 g·kg^{-1} (a) had a slight upward convexity near 800 nm compared to the other curves (b–f). This may be due to its low organic matter content, which does not completely mask the minerals. In the visible wavelength band, the soil

reflectance values in groups (c) and (f) were close, those in groups (c) and (d) were closer, and those in groups (a) and (b) differed significantly. In the near-infrared band, the soil reflectance values differed considerably in all the groups. Soil spectral curves have distinct absorption valleys near 1400, 1900, and 2200 nm. The bands near 1400 and 1900 nm are associated with water vibrations. The band near 2200 nm is related to the intensity of kaolinite with the dioctahedral layer of a mineral structure.

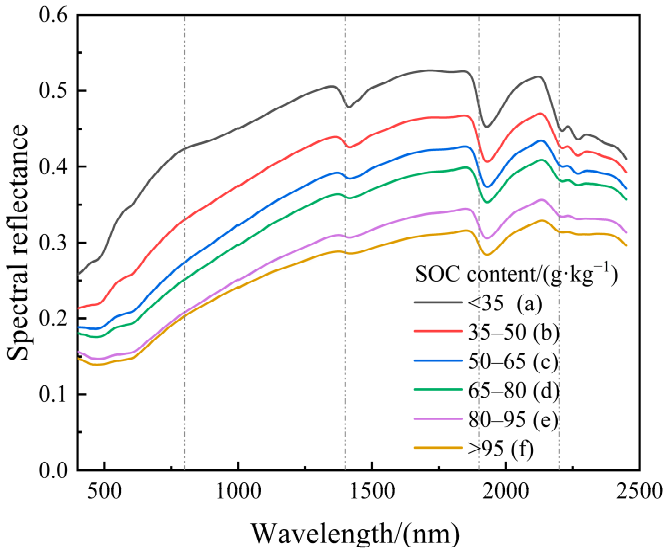

Figure 2. Plot of soil spectra versus soil organic carbon (SOC) content.

3.3. SOC Prediction Accuracy with Strategy I

The modeling results of the SOC content with Strategy I are shown in Figure 3. When using only local-SSL samples as the calibration set, the model based on Set1 (R^2 = 0.83, RMSE = 9.25, RPD = 2.63, RPIQ = 4.60) yielded better predictions than the models based on Set2 (R^2 = 0.76, RMSE = 10.59, RPD = 2.21, RPIQ = 3.29) and Set3 (R^2 = 0.53, RMSE = 14.19, RPD = 1.62, RPIQ = 1.95). In terms of RPD values, the PLSR models achieved excellent, quantitative, and fair predictions with calibration/validation Set1, Set2, and Set3, respectively. The value of RPIQ also decreased gradually. The predicted results for SOC content were related to the proportion of the samples used for calibration, in agreement with previous findings [30]. Similarly, Guerrero et al. found that when the size of the spike subset increased (i.e., having a larger number of samples in the calibration), it had a positive effect on the model prediction accuracy [9].

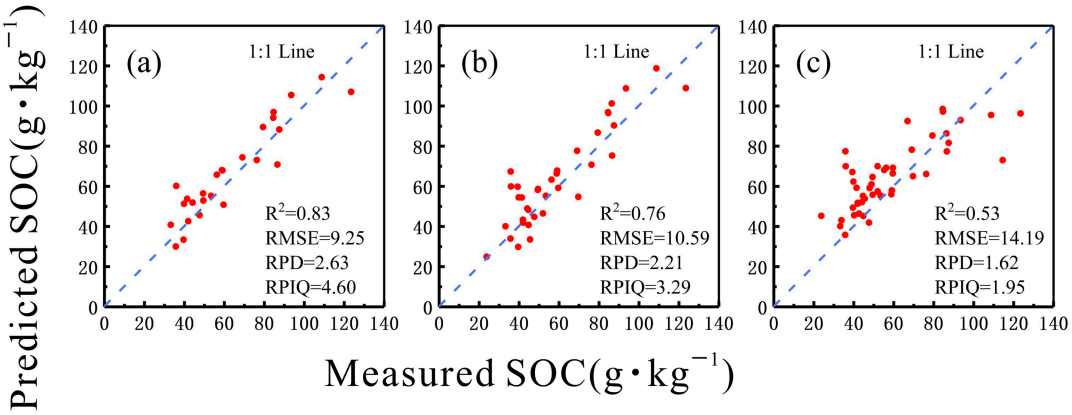

Figure 3. Validation results of Strategy I for soil organic carbon (SOC) prediction. (**a**) Set1 selecting 63 samples from local-SSL. (**b**) Set2 selecting 53 samples from local-SSL. (**c**) Set3 selecting 43 samples from local-SSL. Each plot is marked with a 1:1 line.

3.4. SOC Prediction Accuracy with Strategy II

Figure 4 shows the SOC prediction results from models based on Set1, Set2, and Set3, which contained a calibration subset with *N* randomly selected samples from global-SSL (Strategy II-i). The prediction performance of the models decreased with an increasing *N* value. Moreover, R^2 values of the models based on Set1 (0.72–0.80), Set2 (0.66–0.73), and Set3 (0.55–0.57) decreased sequentially. This trend is similar to that observed for the predictions of Strategy I when using the same *N* values. Excluding Set3, the models based on Set1 and Set2 performed worse for the prediction with Strategy II than with Strategy I. This may be due to the introduction of some samples that were spectrally different from the samples in local-SSL when the calibration subset was randomly selected from global-SSL. The model based on Set3 performed better for prediction with Strategy II than with Strategy I, which may be due to the low ratio of the calibration/validation set (43/46) in Strategy I compared with Strategy II (53/46, 93/46, and 143/46). The predicted SOC content of Set1, Set2, and Set3 all show a tendency for the predicted values to be higher than the measured SOC content, and the overall range of the predicted SOC content is smaller than the measured SOC content. Thus, the model based on the sample number of the calibration set in Strategy I did not perform well in SOC prediction.

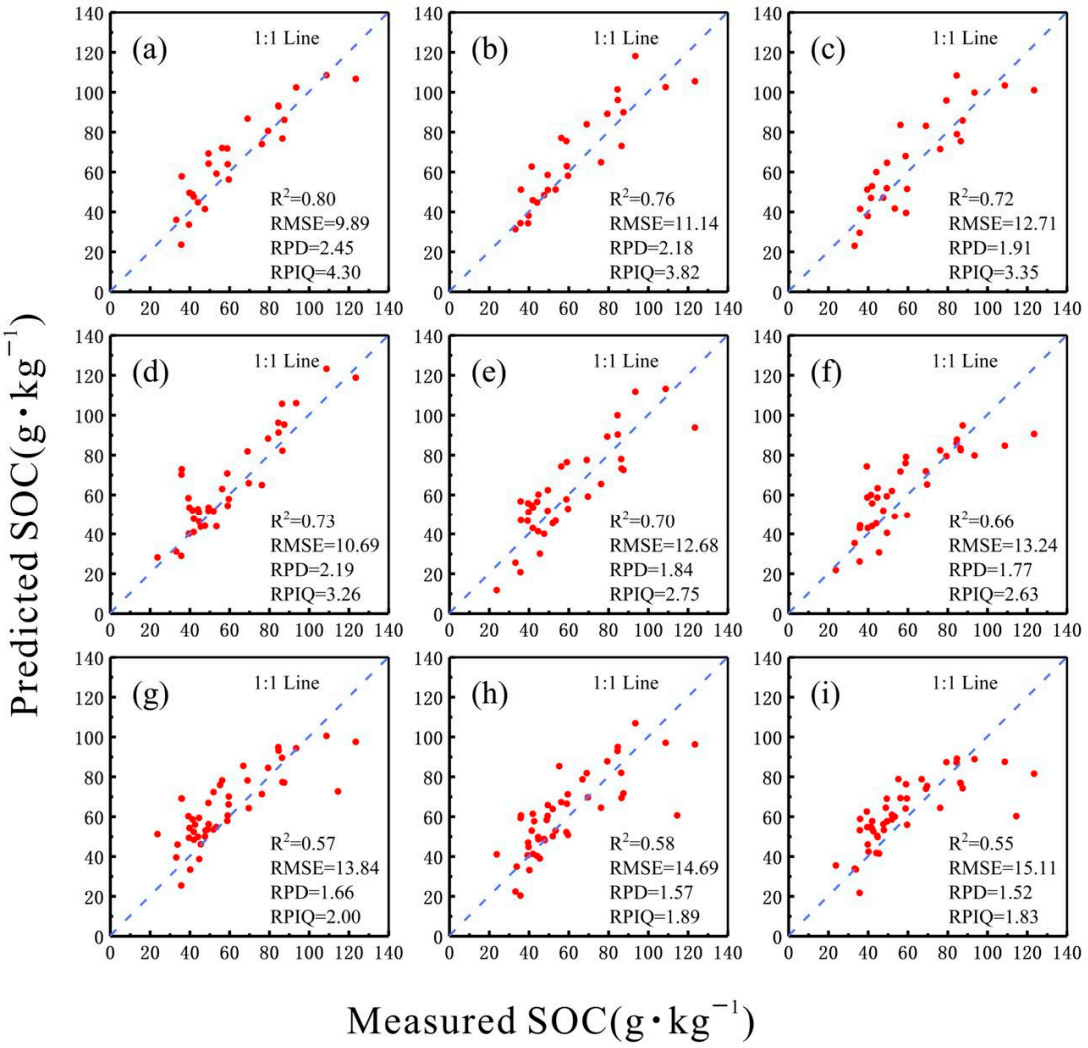

Figure 4. Validation results of strategies II-i for soil organic carbon (SOC) prediction. (**a**–**c**) Set1 with 10, 50, and 100 samples selected from global-SSL, respectively. (**d**–**f**) Set2 with 10, 50, and 100 samples selected from global-SSL, respectively. (**g**–**i**) Set3 with 10, 50, and 100 samples selected from global-SSL, respectively. Each plot is marked with a 1:1 line.

Figures 5 and 6, respectively, depict the validation results of Strategies II-ii and II-iii that randomly select *N* global-SSL samples with low and high spectral similarity to local-SSL samples as the calibration subset. We compared the corresponding results between the two modeling approaches. Excluding Figures 5g and 6g with equal R^2 values (0.58), all cases showed that the modeling approach with Strategy II-iii achieved better predictions than the modeling approach with Strategy II-ii in terms of improved R^2, RMSE, and RPD, RPIQ. In the case of N = 10, the models based on Set1, Set2, and Set3 in Strategy II-iii (Figure 6a,d,g) performed slightly better for prediction than the corresponding models with Strategy II-ii (Figure 5a,d,g). In the case of N = 100, the prediction performance of the models based on Set1, Set2, and Set3 in Strategy II-iii (Figure 6c,f,i) was much better than that of the corresponding models with Strategy II-ii (Figure 5c,f,i).

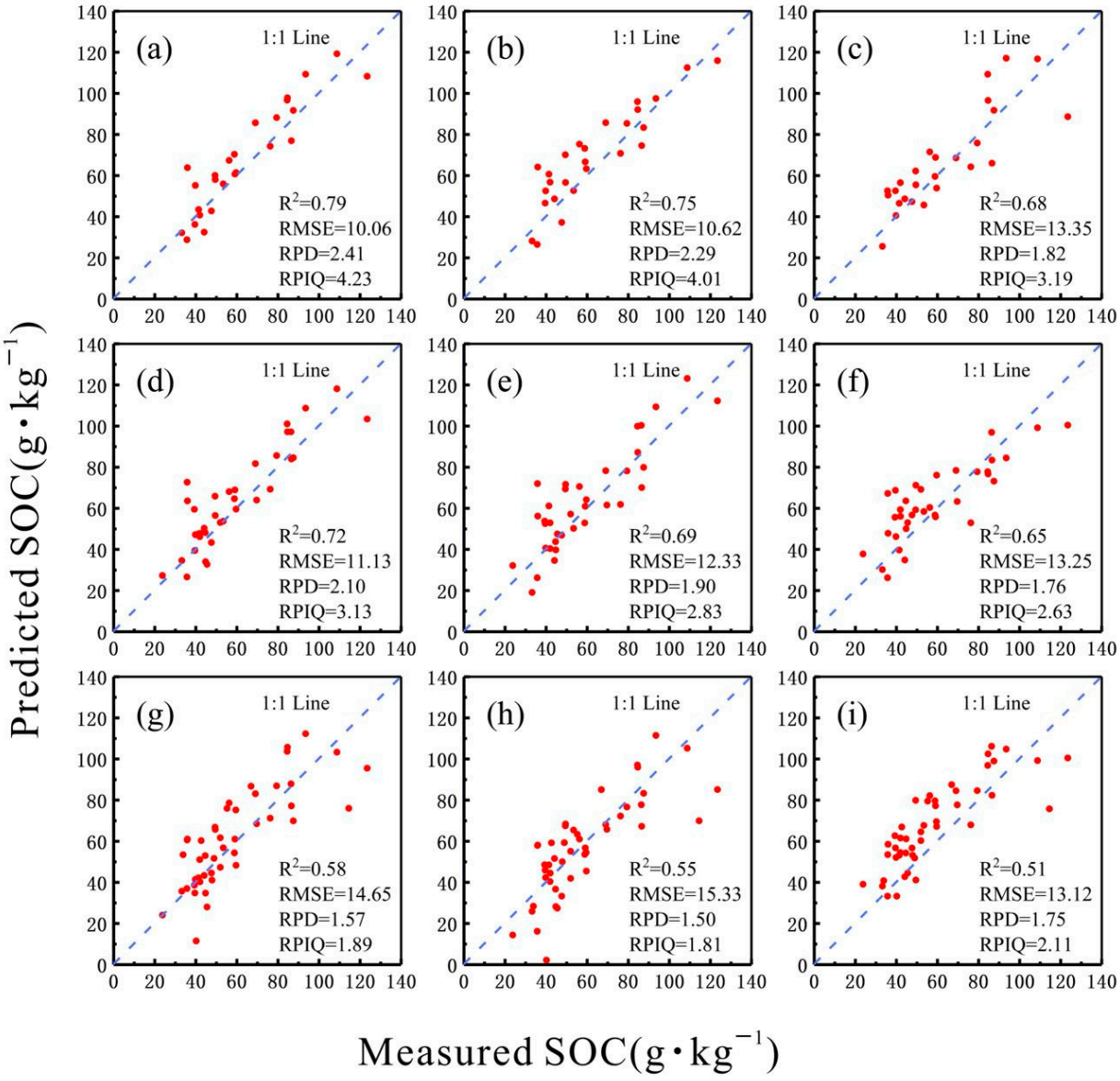

Figure 5. Validation results of strategies II-ii for soil organic carbon (SOC) prediction. (**a–c**) Set1 with 10, 50, and 100 samples selected from global-SSL, respectively. (**d–f**) Set2 with 10, 50, and 100 samples selected from global-SSL, respectively. (**g–i**) Set3 with 10, 50, and 100 samples selected from global-SSL, respectively. Each plot is marked with a 1:1 line.

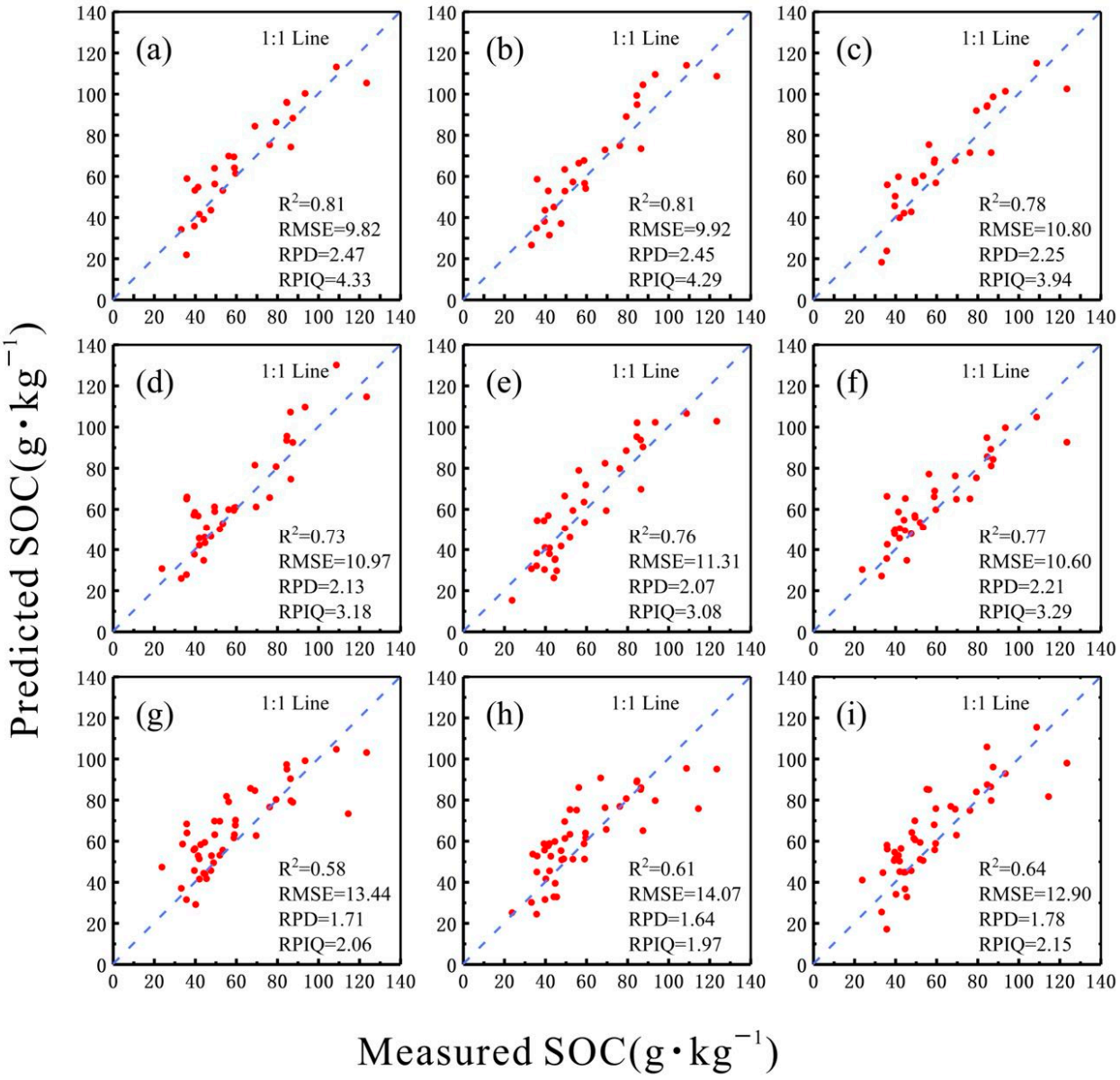

Figure 6. Validation results of strategies II-iii for soil organic carbon (SOC) content prediction. (**a**–**c**) Set1 with 10, 50, and 100 samples selected from global-SSL, respectively. (**d**–**f**) Set2 with 10, 50, and 100 samples selected from global-SSL, respectively. (**g**–**i**) Set3 with 10, 50, and 100 samples selected from global-SSL, respectively. Each plot is marked with a 1:1 line.

The dispersion of the predicted SOC values obtained by modeling with Strategy II-iii was more pronounced than the results of the other two modeling approaches with Strategies II-i and II-ii; R^2 whose values were lower for the latter two approaches than for the former approach (Figures 4–6). Modeling with high-similarity samples produced better prediction results than modeling with randomly selected samples. Accordingly, selecting samples with high spectral similarity should be undertaken as the calibration subset contributes to the improved modeling accuracy of SOC. Similar results were reported by Viscarra Rossel et al., who found that for some attributes (e.g., SOC, pH, clay, and sand content), modeling accuracy was improved overall by classifying spectral similarities from the global spectral library and then modeling them separately [15]. Araújo et al. found that dividing the global dataset into more mineralogically uniform clusters, regardless of geographical origin, effectively improved the prediction [22]. However, in these instances, excluding Set3, modeling with Strategy II did not improve the prediction performance

compared with Strategy I. This observation may be due to the fact that the high spectral similarity is based on an overall evaluation of the entire global-SSL versus local-SSL. It is possible that global-SSL samples with moderately high spectral similarity to local-SSL samples were selected in the actual random selection, while all high-similarity samples were discarded, resulting in a loss of useful information. There is also a possibility that only certain global-SSL samples with high spectral similarity to some local-SSL samples were selected so that part of the valid information was masked.

The modeling results obtained with Strategy II were based on the mean of 100 replicates. It is, therefore, reasonable to conclude that although the post-mean results do not show the advantage of using high-similarity samples selected from global-SSL as a reference calibration set, there must be cases where the predicted results are greater than the mean. Similarly, Guerrero et al. [9] compared the predictive accuracy of PLSR models for SOC content before and after 'spiking', finding that in three of the four target sites, the predictive accuracy improved after 'spiking '. Based on these results, global-SSL samples with high spectral similarity to local-SSL samples were selected as a calibration subset in Strategy III. We, therefore, conjecture that the modeling approach with Strategy III is responsible for prediction accuracy improvement.

3.5. SOC Prediction Accuracy with Strategy III

The prediction results of the two methods used in Strategy III are shown in Figures 7 and 8. The *N*th point represents the result of selecting the top *N* spectral samples from global-SSL that were the most similar to each local-SSL sample as the calibration subset, and the *x*-axis coordinate represents the number of reference global-SSL samples. Given the small number of local-SSL samples, we only considered the case of selecting <100 reference samples. R^2, RMSE, RPD, and RPIQ values did not show clear patterns with an increasing *N* value, and the reasons have been explained in detail in Section 2.5.

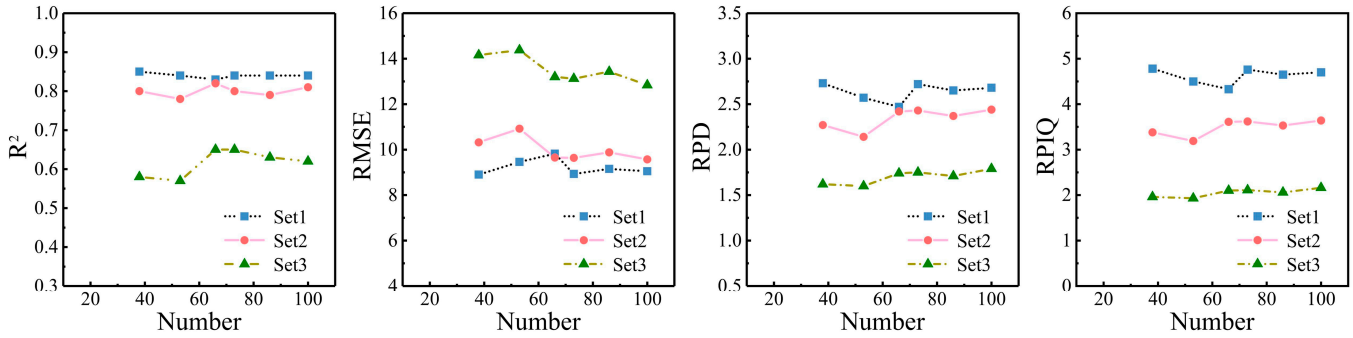

Figure 7. Validation results of Strategy III-i for soil organic carbon (SOC) prediction.

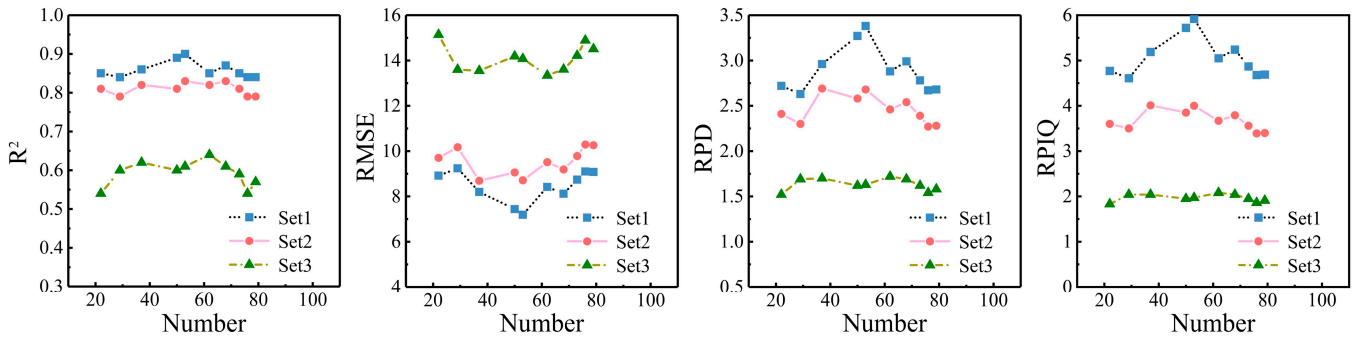

Figure 8. Validation results of Strategy III-ii for soil organic carbon (SOC) prediction.

The best validation results were obtained from the model with Strategy III-i when ~70 global-SSL samples were selected as the calibration subset. The model with Strategy

III-ii produced the best validation results when ~50 global-SSL samples were selected as the calibration subset. Wetterlind et al. examined the performance of a small farm-scale calibration (25 samples) versus a larger national soil library (396 samples) for within-field soil characterization based on vis-NIR spectroscopy. They tested whether site-specific sample sets selected from the national library, including a spectral library consisting of the 50 samples most similar to the local-site samples, could improve the prediction performance. It was found that trends in the clay and SOC content were better predicted using a reduced national soil library compared with using the entire national library. In some cases, the model even outperformed predictions using local calibration [27].

The model with Strategy III-ii also yielded better validation results than the model with strategy III-i when the same number of samples were selected as the reference calibration set. Compared with Strategies I and II, the model validation results were improved for both methods selecting the calibration subset in Strategy III. In the best case, the prediction with Strategy III-ii achieved the highest accuracy: Set1 (R^2 = 0.90, RMSE = 7.19, RPD = 3.38), Set2 (R^2 = 0.83, RMSE = 8.71, RPD = 2.68), and Set3 (R^2 = 0.64, RMSE = 13.34, RPD = 1.72). Compared with Strategy I, Strategy III-ii improved R^2 by 0.07, 0.07, 0.11, decreased RMSE by 2.06, 1.88, 0.85, and improved RPD by 0.75, 0.47, 0.1, for Set1, Set2, and Set3, respectively.

Overall, our results suggest that the use of soil samples with similar properties from the target area facilitates accurate SOC predictions due to the similarities in environmental factors, such as geographic settings and parent material [49]. However, we found that the model performed very well when the calibration set was established with a calibration subset from global-SSL (with top *N* spectral samples similar to each local sample) and the 'spiking' set from local-SSL (generally RPD > 2.5). Thus, it is a feasible modeling method to improve SOC prediction and reduce the sample number of the 'spiking' set from local-SSL.

4. Conclusions

This study compared the performance of three different modeling strategies for soil organic carbon (SOC) prediction in target forest areas based on local and global soil spectral libraries (local-SSL and global-SSL, respectively). These strategies used a calibration set consisting of (i) only local-SSL samples, (ii) a randomly selected calibration subset from global-SSL and the 'spiking' set from local-SSL, and (iii) a deliberately selected calibration subset from global-SSL and the 'spiking' set from local-SSL.

We found that when randomly selecting calibration samples from global-SSL combined with a small set of 'spiking' samples from local-SSL, this did not necessarily achieve consistently better SOC predictions. By contrast, the use of spectrally similar global-SSL samples provided better predictions. Using the optimal modeling strategy, the top *N* spectral samples with high similarity to each local sample were extracted from global-SSL as the calibration subset. This calibration subset from global-SSL and the 'spiking' set from local-SSL were used as the calibration set, which improved the prediction results in the target area.

We suggest that when selecting spectrally similar samples from global-SSL, each of the samples in local-SSL should be taken into account without losing sight of the others and favoring a few specific samples. This strategy is inexpensive and beneficial when expert knowledge about soil classification is lacking. If one cannot afford the high cost of laborious testing or intensive sampling in mountainous forest areas with challenging terrains, our proposed modeling approach, combined with an existing soil spectral library, is an effective way to obtain sufficiently reliable SOC data. Determining the optimal sample number for the reference calibration sets selected from the soil spectral library and identifying them within it should be carried out in future work.

Author Contributions: Conceptualization, M.L.; data curation, M.L. and Z.X.; formal analysis, M.L. and J.G.; methodology, M.L., X.G., and X.Z.; software, M.L.; validation, M.L., T.Y., and X.Z.; investigation, Z.X. and J.L.; writing—original draft preparation, M.L.; writing—review and editing, M.L., J.G., and Z.X.; supervision, X.Z.; project administration, X.Z.; funding acquisition, X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Chinese national key project of high-resolution Earth observation system, grant number 82-Y50G22-9001-22/23.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data can be acquired at https://www.isric.org/data/ (accessed on 14 June 2022).

Acknowledgments: The authors would like to thank anonymous reviewers for their valuable comments and remarks.

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

References

1. Lal, R. Forest soils and carbon sequestration. *Forest Ecol. Manag.* **2005**, *220*, 242–258. [CrossRef]
2. Lal, R. Soil carbon sequestration to mitigate climate change. *Geoderma* **2004**, *123*, 1–22. [CrossRef]
3. Rossel, R.A.V.; 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]
4. Shi, Z.; Ji, W.; Rossel, R.A.V.; 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]
5. Chen, Y.; Wang, J.; Liu, G.; Yang, Y.; Liu, Z.; Deng, H. Hyperspectral Estimation Model of Forest Soil Organic Matter in Northwest Yunnan Province, China. *Forests* **2019**, *10*, 217. [CrossRef]
6. Liu, S.; Shen, H.; Chen, S.; Zhao, X.; Fang, J. Estimating forest soil organic carbon content using vis-NIR spectroscopy: Implications for large-scale soil carbon spectroscopic assessment. *Geoderma* **2019**, *348*, 37–44. [CrossRef]
7. Yang, H.; Li, J. Predictions of soil organic carbon using laboratory-based hyperspectral data in the northern Tianshan mountains, China. *Environ. Monit. Assess.* **2013**, *185*, 3897–3908. [CrossRef]
8. Reda, R.; Saffaj, T.; Ilham, B.; Saidi, O.; Hadrami, E. A comparative study between a new method and other machine learning algorithms for soil organic carbon and total nitrogen prediction using near infrared spectroscopy. *Chemometr. Intell. Lab.* **2019**, *195*, 103873. [CrossRef]
9. Guerrero, C.; Stenberg, B.; Wetterlind, J.; Rossel, R.; Sinoga, J. 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]
10. Stevens, A.; Nocita, M.; Toth, 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]
11. Hill, J.; Udelhoven, T.; Vohland, M.; Stevens, A. *The Use of Laboratory Spectroscopy and Optical Remote Sensing for Estimating Soil Properties*; Springer: Dordrecht, The Netherlands, 2010; pp. 67–85.
12. Yaolin, L.; Qinghu, J.; Teng, F.; Junjie, W.; Tiezhu, S.; Kai, G.; Xiran, L.; Yiyun, C. Transferability of a Visible and Near-Infrared Model for Soil Organic Matter Estimation in Riparian Landscapes. *Remote Sens.* **2014**, *6*, 4305–4322.
13. Liu, Y.; Shi, Z.; Zhang, G.; Chen, Y.; Li, S.; Hong, Y.; Shi, T.; Wang, J.; Liu, Y. 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. *Remote Sens.* **2018**, *10*, 1747. [CrossRef]
14. Castaldi, F.; Chabrillat, S.; Chartin, C.; Genot, V.; Jones, A.R.; 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]
15. Viscarra Rossel, R.A.; Behrens, T.; Ben-Dor, E.; Brown, D.; Demattê, J.; Shepherd, K.; Shi, Z.; Stenberg, B.; Stevens, A.; Adamchuk, V. A global spectral library to characterize the world's soil. *Earth-Sci. Rev.* **2016**, *155*, 198–230. [CrossRef]
16. 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]
17. Guerrero, C.; Wetterlind, J.; Stenberg, B.; Mouazen, A.M.; Gabarrón-Galeote, M.A.; RuizSinoga, J.D.; Zornoza, R.; Viscarra Rossel, R.A. Do we really need large spectral libraries for local scale SOC assessment with NIR spectroscopy? *Soil Tillage Res.* **2015**, *155*, 501–509. [CrossRef]
18. 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]
19. Rossel, R.A.V.; Webster, R. Predicting soil properties from the Australian soil visible–near infrared spectroscopic database. *Eur. J. Soil Sci.* **2012**, *63*, 848–860. [CrossRef]
20. Seidel, M.; Hutengs, C.; Ludwig, B.; Sren, T.B.; Vohland, M. Strategies for the efficient estimation of soil organic carbon at the field scale with vis-NIR spectroscopy: Spectral libraries and spiking vs. local calibrations. *Geoderma* **2019**, *354*, 113856. [CrossRef]
21. Xiaomi, W.; Yiyun, C.; Long, G.; Leilei, 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.

22. Araújo, S.R.; Wetterlind, J.; Demattê, J.A.M.; Stenberg, B. Improving the prediction performance of a large tropical vis-NIR spectroscopic soil library from Brazil by clustering into smaller subsets or use of data mining calibration techniques. *Eur. J. Soil Sci.* **2014**, *65*, 718–729. [CrossRef]
23. 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]
24. 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 Spec.* **2010**, *18*, 167–176. [CrossRef]
25. Guy, A.L.; Siciliano, S.D.; Lamb, E.G. Spiking regional vis-NIR calibration models with local samples to predict soil organic carbon in two High Arctic polar deserts using a vis-NIR probe. *Can. J. Soil Sci.* **2015**, *95*, 237–249. [CrossRef]
26. 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]
27. 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]
28. 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]
29. Kuang, B.; Mouazen, A.M. Influence of number of samples on prediction error of visible and near infrared spectroscopy of selected soil properties at farm scale. *Eur. J. Soil Sci.* **2012**, *1*, 421–429. [CrossRef]
30. Shengxiang, X.; Xuezheng, S.; Meiyan, W.; Yongcun, Z.; Jingdong, M. Effects of Subsetting by Parent Materials on Prediction of Soil Organic Matter Content in a Hilly Area Using Vis–NIR Spectroscopy. *PLoS ONE* **2016**, *11*, e151536.
31. Ramirez-Lopez, L.; Behrens, T.; Schmidt, K.; Rossel, R.A.V.; Demattê, J.A.M.; Scholten, T. Distance and similarity-search metrics for use with soil vis–NIR spectra. *Geoderma* **2013**, *199*, 43–53. [CrossRef]
32. Farifteh, J.; Meer, F.V.D.; Carranza, E.J.M. Similarity measures for spectral discrimination of salt-affected soils. *Int. J. Remote Sens.* **2007**, *28*, 5273–5293. [CrossRef]
33. Lugassi, R.; Ben-Dor, E.; Eshel, G. A spectral-based method for reconstructing spatial distributions of soil surface temperature during simulated fire events. *Remote Sens. Environ.* **2010**, *114*, 322–331. [CrossRef]
34. Changlong, W.; Yuguo, Z.; Decheng, L.; Ganlin, Z.; Dengwei, W.; Jike, C. Prediction of soil organic matter and cation exchange capacity based on spectral similarity measuring. *Trans. Chin. Soc. Agric. Eng. Trans. CSAE* **2014**, *1*, 81–88.
35. Hongda, L.; Decheng, L.; Rong, Z. Estimation of Soil Organic Carbon Based on Spectral Similarity Matching. *Acta Pedol. Sin.* **2021**, *58*, 1224–1233.
36. Kruse, F.A.; Lefkoff, A.B.; Boardman, J.W.; Heidebrecht, K.B.; Shapiro, A.T.; Barloon, P.J.; Goetz, A. The Spectral Image Processing System (SIPS)-Interactive Visualization and Analysis of Imaging Spectrometer Data. *Remote Sens. Environ.* **1993**, *44*, 145–163. [CrossRef]
37. Park, B.; Windham, W.R.; Lawrence, K.C.; Smith, D.P. Contaminant Classification of Poultry Hyperspectral Imagery using a Spectral Angle Mapper Algorithm. *Biosyst. Eng.* **2007**, *96*, 323–333. [CrossRef]
38. 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]
39. 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]
40. Wold, S.; Sj Str, M.M.; Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130. [CrossRef]
41. Cambou, A.; Barthès, B.G.; Moulin, P.; Chauvin, L.; Faye, E.H.; Masse, D.; Chevallier, T.; Chapuis-Lardy, L. Prediction of soil carbon and nitrogen contents using visible and near infrared diffuse reflectance spectroscopy in varying salt-affected soils in Sine Saloum (Senegal). *Catena* **2022**, *212*, 106075. [CrossRef]
42. Guo, P.; Li, T.; Gao, H.; Chen, X.; Cui, Y.; Huang, Y. Evaluating Calibration and Spectral Variable Selection Methods for Predicting Three Soil Nutrients Using Vis-NIR Spectroscopy. *Remote Sens.* **2021**, *13*, 4000. [CrossRef]
43. Ba, Y.; Liu, J.; Han, J.; Zhang, X. Application of Vis-NIR spectroscopy for determination the content of organic matter in saline-alkali soils. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2020**, *229*, 117863. [CrossRef] [PubMed]
44. Chen, S.; Xu, D.; Li, S.; Ji, W.; Yang, M.; Zhou, Y.; Hu, M.; Xu, H.; Shi, Z. Monitoring soil organic carbon in alpine soils using in situ vis km IR spectroscopy and a multilayer perceptron. *Land Degrad. Dev.* **2020**, *31*, 1026–1038. [CrossRef]
45. Kennard, R.W.; Stone, L.A. Computer aided design of experiments. *Technometrics* **1969**, *11*, 137–148. [CrossRef]
46. Gholizadeh, A.; Rossel, R.A.V.; Saberioon, M.; Boruvka, L.; Kratina, J.; Pavlu, L. National-scale spectroscopic assessment of soil organic carbon in forests of the Czech Republic. *Geoderma* **2021**, *385*, 114832. [CrossRef]
47. Jaconi, A.; Don, A.; Freibauer, A. Prediction of soil organic carbon at the country scale: Stratification strategies for near-infrared data. *Eur. J. Soil Sci.* **2017**, *68*, 919–929. [CrossRef]
48. Rossel, R.A.V.; McGlynn, R.N.; McBratney, A.B. Determining the composition of mineral-organic mixes using UV–vis–NIR diffuse reflectance spectroscopy—ScienceDirect. *Geoderma* **2006**, *137*, 70–82. [CrossRef]

49. Savvides, A.; Corstanje, R.; Baxter, S.J.; Rawlins, B.G.; Lark, R.M. The relationship between diffuse spectral reflectance of the soil and its cation exchange capacity is scale-dependent. *Geoderma* **2010**, *154*, 353–358. [CrossRef]
50. Stoner, E.R.; Baumgardner, M.F. Characteristic Variations in Reflectance of Surface Soils. *Soil Sci. Soc. Am. J.* **1982**, *45*, 1161–1165. [CrossRef]