land

MDPI

Article

Development of Soil Fertility Index Using Machine Learning and Visible-Near-Infrared Spectroscopy

Xiaolin Jia [1], Yi Fang [1], Bifeng Hu [2], Baobao Yu [1] and Yin Zhou [3,*]

[1] College of Surveying and Geo-Informatics, North China University of Water Resources and Electric Power, Zhengzhou 450046, China; jiaxiaolin@ncwu.edu.cn (X.J.); 202004905@stu.ncwu.edu.cn (Y.F.); x20211151015@stu.ncwu.edu.cn (B.Y.)
[2] Department of Land Resource Management, School of Public Finance and Public Administration, Jiangxi University of Finance and Economics, Nanchang 330013, China; hubifeng@zju.edu.cn
[3] Institute of Land and Urban-Rural Development, Zhejiang University of Finance and Economics, Hangzhou 310018, China
* Correspondence: zhouyin@zju.edu.cn

Abstract: An accurate assessment of soil fertility is crucial for monitoring environmental dynamics, improving agricultural productivity, and achieving sustainable land management and utilization. The inherent complexity and spatiotemporal heterogeneity of soils result in significant challenges in soil fertility assessment. Therefore, this study focused on developing a rapid, economical, and precise approach to evaluate soil fertility through the application of visible-near-infrared spectroscopy (VNIR). To achieve this, we utilized the Land Use and Cover Area Frame Survey (LUCAS) dataset and employed a variety of prediction models, including partial least squares regression, support vector machines (SVMs), random forest, and convolutional neural networks, to estimate various soil properties and overall soil fertility. The results showed that the SVM model had the highest prediction accuracy, particularly for clay content (coefficient of determination (R^2) = 0.79, ratio of performance to interquartile range (RPIQ) = 3.04), pH (R^2 = 0.84, RPIQ = 4.54), total nitrogen (N) (R^2 = 0.80, RPIQ = 2.40), and cation exchange capacity (CEC) (R^2 = 0.83, RPIQ = 3.16). A soil fertility index (SFI) was developed based on factor analysis, integrating nine essential soil properties: clay content, silt content, sand content, pH, carbonate content, N, soluble phosphorus, soluble potassium, and CEC. We compared direct and indirect prediction models for estimating SFI and found that both models showed high accuracy (mean value of R^2 = 0.80, mean value of RPIQ = 2.21). Additionally, SFI was classified into five classes to provide insights for precision agriculture. The kappa coefficient was 0.63, which indicated that the SFI evaluation results between VNIR and chemical analysis were relatively consistent. This study provides a theoretical foundation of real-time soil fertility monitoring for the optimization of agricultural practices.

Keywords: soil fertility; VNIR; machine learning; precision agriculture; land management

Citation: Jia, X.; Fang, Y.; Hu, B.; Yu, B.; Zhou, Y. Development of Soil Fertility Index Using Machine Learning and Visible-Near-Infrared Spectroscopy. *Land* **2023**, *12*, 2155. https://doi.org/10.3390/land12122155

Academic Editors: Shengbiao Wu, Xiao Zhang and Xidong Chen

Received: 30 October 2023
Revised: 9 December 2023
Accepted: 11 December 2023
Published: 12 December 2023

1. Introduction

Soil fertility is defined as the ability of the soil to provide available nutrients for crop production [1]. Not only is it fundamental for sustainable crop production and agricultural development but it also plays an important role in maintaining ecosystem health [2]. Deficiencies in soil nutrients can severely reduce crop yields, whereas excessive soil nutrients can decrease production profits and result in negative environmental impacts [3]. However, due to the high spatiotemporal heterogeneity of soil, the management of soil fertility is a great challenge. Assessing soil fertility accurately is crucial for monitoring soil environment dynamics and improving agricultural productivity [4].

The soil fertility index (SFI), a weighted combination of various soil properties, is generally used to assess soil fertility [2,5]. It is usually costly to obtain soil information

via traditional chemical analysis, especially for multiple soil properties [6]. Proximal soil-sensing technologies, particularly visible-near-infrared spectroscopy (VNIR), provide a fast, economical, and accurate approach to soil property analysis. A number of soil properties can be estimated using prediction models based on the detailed spectral absorption and reflection characteristics of substances [7,8]. These predictable properties can often be used for SFI development, including soil properties with direct spectral responses in the VNIR spectra (such as moisture content and clay content), as well as secondary soil properties that may covary with the main soil properties (such as pH) [9,10]. Therefore, proximal soil sensing technology has great potential for assessing soil fertility and estimating SFI.

VNIR has been widely used in the estimation of SFI. Viscarra Rossel et al. [11] developed an SFI with clay content, base saturation, cation exchange capacity (CEC), and organic matter (OM) at the field scale. They successfully predicted three categories of soil fertility, with accuracy ranging from 61% to 75%, by integrating VNIR with topographic data and employing a decision tree model. Askari et al. [12] applied VNIR and partial least squares regression (PLSR) models to predict specific soil characteristics and soil quality indices for grassland and arable land. The prediction results exhibited a high coefficient of determination (R^2) of 0.89 for grassland soil quality (indicators included organic carbon (OC), carbon-to-nitrogen ratio, and bulk density) and 0.81 for cropland (indicators included aggregate size distribution, bulk density, carbon-to-nitrogen ration, extractable magnesium, total nitrogen (N), penetration resistance, and respiration). Yang et al. [13] estimated the SFI (indicators including pH, OM, N, available phosphorus (P), available potassium (K), CEC, texture, available nitrogen, total phosphorus, and total potassium) of paddy fields in southern China using the VNIR and PLSR models, achieving an R^2 of 0.80 and a ratio of performance to interquartile range (RPIQ) of 3.12. Munnaf and Mouazen [14] implemented soil scanning using an online VNIR sensor (CompactSpec, Tec5 Technology, Steinbach, Germany) and calibrated the SFI (indicators included pH, OC, P, K, available magnesium, available sodium, moisture content) model using the PLSR algorithm, resulting in a robust prediction with an R^2 of 0.75 and a ratio of prediction to deviation (RPD) of 2.01. Based on these studies, the construction of SFI usually integrates different soil properties without a recognized combination, leading to a lack of comparability among different results. Furthermore, most studies have focused on the construction of SFI, without considering the impact of soil property prediction accuracy.

To address these problems, this study applied different models to predict properties of soil fertility with spectroscopy, and the model performances were compared. The SFI was estimated directly from VNIR and through soil properties predicted from spectroscopy, respectively. This study is intended to develop an economical and efficient method to evaluate SFI, which may provide decision makers with effective information for soil fertility monitoring and management.

2. Materials and Methods

2.1. Data Collection

The Land Use and Cover Area Frame Survey (LUCAS) published information on 21,782 topsoil samples (0–20 cm) from 28 European Union member states in November 2020. In this study, we utilized French soil samples (2792) from the LUCAS dataset to establish spectral prediction models for soil properties and SFI (Figure 1). Approximately 55% of the samples were collected in cropland, with the rest obtained from woodland, shrubland, and grassland. After removing vegetation residue and litter, air-drying, grinding, and sieving (pore size < 2 mm), the samples were shipped for chemical analysis, including particle size distribution (clay content, silt content, and sand content), pH, OC, carbonate content ($CaCO_3$), N, P, K, CEC, and multispectral reflectance [15]. Soil properties were determined using international standard methods. Multispectral reflectance was measured using a FOSS XDS spectrometer (FOSS, Hillerød, Denmark), with a wavelength range of 400–500 nm and an interval of 0.5 nm. In this study, Haar wavelet transformation was used on the preprocessed spectra to reduce noise and enhance the features.

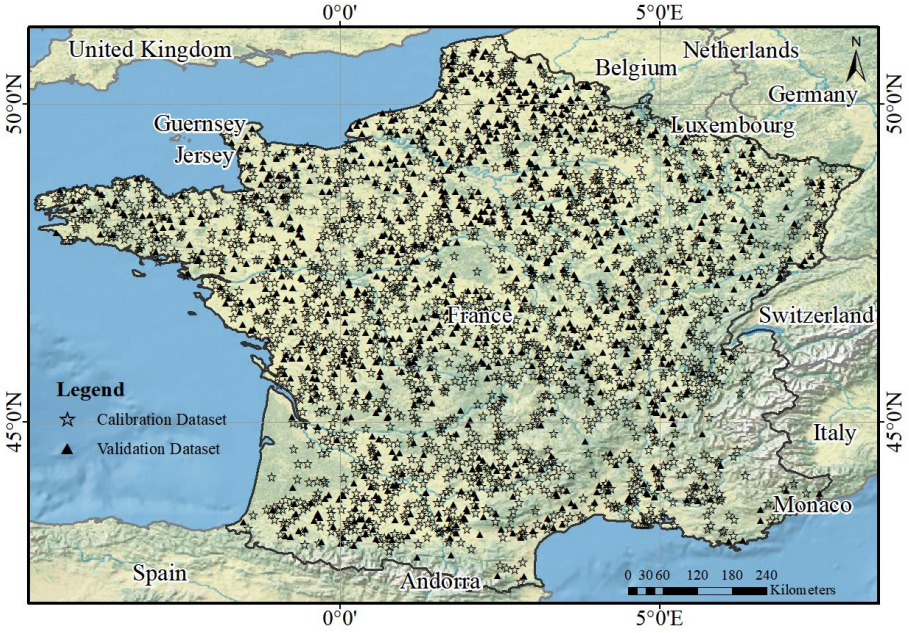

Figure 1. Location of the soil sampling sites.

2.2. Soil Fertility Index

The SFI is an indicator integrating various soil physicochemical properties for the assessment of soil health from the perspectives of environmental management and soil production potential [16]. The derivation of the SFI comprises three steps: (i) selection of indices, (ii) calculation of index weights, and (iii) calculation of the comprehensive index. First, Pearson correlation analysis was performed for the 10 soil properties, and variables that exhibited pairwise correlations higher than 0.85 were excluded to minimize collinearity among the variables. Then, factor analysis was used to select the soil properties with linear combination coefficients exceeding 0.05 for each principal component. Considering the intrinsic relationships among the soil properties, factor analysis was used to calculate the weights of each index. The calculation steps involved (i) determining the linear combination coefficients of each soil property in different principal components using Equation (1); (ii) determining the comprehensive score coefficient of each soil property using Equation (2); and (iii) calculating the weights of each index by normalizing the comprehensive score coefficients [17]. Finally, the SFI result was calculated through weighted summation and transformed into a range between 0 and 1 and then was classified into different levels based on management practices.

$$LCC_{nm} = \frac{CL_{nm}}{\sqrt[2]{q_m}} \quad (1)$$

$$CSC_n = \frac{\sum_m LCC_{nm} \times PVE_m}{CP} \quad (2)$$

where LCC_{nm} is the linear combination coefficient of soil property n in principal component m; CL_{nm} is the rotated component loading of soil property n in principal component m; q_m is the rotated characteristic root of principal component m; CSC_n is the comprehensive score coefficient of soil property n; PVE_m is the rotated explained variance proportion of principal component m; and CP is the cumulative explained variance proportion.

2.3. Spectral Modeling

In this study, the predictive capabilities for various soil properties of different models were compared based on VNIR, including linear, nonlinear, and machine learning models. Then, we established relationship models between soil fertility and VNIR through indirect and direct prediction methods, respectively. The Kennard–Stone (KS) algorithm was

utilized to select the calibration and validation datasets [18]. Two-thirds of the soil samples with significant spectral differences were selected as the calibration dataset by calculating the Euclidean distances between samples. Meanwhile, the remaining samples were used for the validation dataset. The workflow of this study is shown in Figure 2.

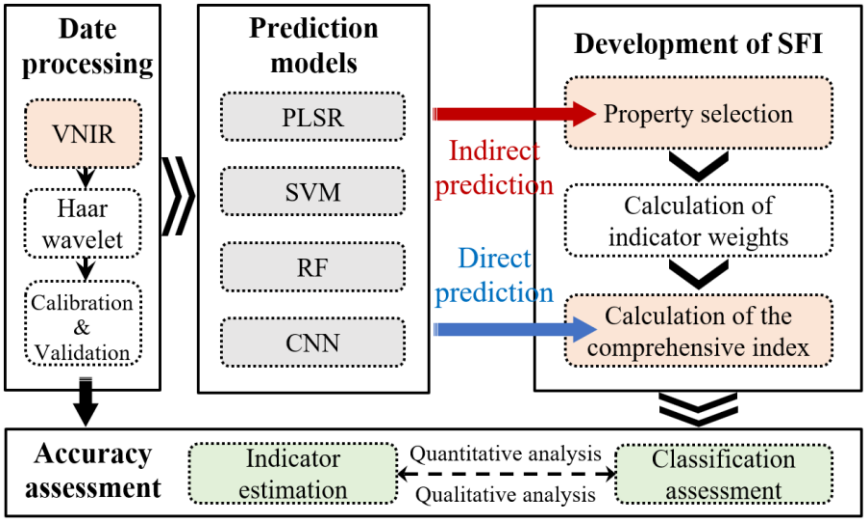

Figure 2. Workflow of the SFI estimation. (VNIR: visible-near-infrared spectra; SVM: support vector machine; RF: random forest; CNN: convolutional neural network).

This study used two methods for estimating SFI using spectral data: (i) an indirect prediction model, where soil properties were predicted using VNIR and SFI was calculated based on these predictions, and (ii) a direct prediction model, which directly predicted SFI using VNIR. We used four different types of models, i.e., PLSR, support vector machine (SVM), random forest (RF), and convolutional neural network (CNN), to construct spectral prediction models. Owing to the large volume of spectral data, PLSR was used to compress the spectral data, enhancing the computational efficiency and prediction accuracy of the SVM, RF, and CNN.

PLSR is a statistical learning method that decomposes the covariance matrix between the dependent and independent variables and transforms the prediction problem into a series of linear regression problems [8,19]. Through PCA and regression analysis, PLSR can effectively handle high-dimensional data and multicollinearity issues but is sensitive to outliers and performs poorly when dealing with nonlinear relationships.

SVM is a linear classifier that projects data into a higher dimensional space using a kernel function and then seeks an optimal hyperplane that maximizes the margin between classes [20,21]. Despite its ability to handle linearly inseparable data, SVM training on large-scale datasets is often associated with a relatively slow computational speed. In addition, the interpretability of the SVM predictions is comparatively limited.

RF is a supervised machine learning method that builds multiple decision trees by randomly selecting features and samples [22,23]. The final prediction result is determined by the average or majority vote of all the decision trees. RF has significant advantages in dealing with problems such as overfitting, multicollinearity, and missing and imbalanced data. However, replicating the model's results is challenging because of the inherent randomness.

CNN is a deep learning algorithm that is built by combining fully connected layers, convolutional layers, and pooling layers. It can automatically extract features from input data and can be trained using a backpropagation algorithm [24]. CNN has powerful expressive capabilities and can learn both local and global features; however, it requires a large amount of training data and computational resources.

2.4. Accuracy Assessment

For the regression models, R^2, root mean square error (RMSE), RPD, and RPIQ of the validation dataset were used to assess the prediction accuracy of soil properties and soil fertility. Generally, lower RMSE and higher R^2, RPD, and RPIQ values lead to higher prediction accuracy [25]. Considering the RPD values, prediction accuracy is often classified as excellent (RPD > 2.5), very good (RPD: 2.0–2.5), good (RPD: 1.8–2.0), fair (RPD: 1.4–1.8), poor (RPD: 1.0–1.4), or very poor (RPD < 1.0).

For classification, the results were evaluated using Cohen's kappa coefficient (k), where a higher k value indicates better classification. According to k values, classification accuracy can be divided into five classes: almost perfect agreement (k > 0.8), substantial agreement (k: 0.6–0.8), moderate agreement (k: 0.4–0.6), fair agreement (k: 0.2–0.4), and slight agreement (k < 0.2) [26].

The factor analysis and KS algorithm, along with the RF and CNN prediction models, were implemented using R software, version 4.3.2 [27]. The preprocessing of spectra and the PLSR and SVM prediction models were conducted using MatLab version 7 (MathWorks Inc., Natick, MA, USA) and PLS_Toolbox 8.5 (Eigenvector Research Inc., Wenatchee, WA, USA).

3. Results

3.1. Summary of the Characteristics of Soil Properties

A statistical summary of the soil properties is presented in Table 1. It has been reported that a wider range of soil properties is beneficial for developing prediction models that cover a substantial SFI range [13]. The coefficient of variation (CV) was used to indicate the degree of variation in soil properties and can be divided into four degrees: exceptionally high variability (CV > 100%), high variability (CV: 50–100%), moderate variability (CV: 20–50%), and low variability (CV < 20%) [28]. The CVs for the soil properties, in descending order of variability, were as follows: $CaCO_3$, P, OC, K, sand content, CEC, N, clay content, silt content, and pH. The exceptionally high variability in $CaCO_3$ (222%) suggests that its distribution is heavily influenced by both the parent material and the external environmental conditions.

Table 1. Statistical summary of the 10 soil properties.

Property	Unit	Min	Mean	Median	Max	SD	CV
clay content	%	2.00	23.30	21.00	77.00	11.07	48%
silt content	%	1.00	47.45	47.00	88.00	17.85	38%
sand content	%	1.00	29.24	24.00	96.00	21.24	73%
pH	-	3.51	6.65	6.76	8.90	1.06	16%
OC	g kg^{-1}	0.00	25.70	19.90	191.50	19.88	77%
$CaCO_3$	g kg^{-1}	0.00	69.38	1.00	944.00	153.88	222%
N	g kg^{-1}	0.20	2.32	1.90	14.00	1.40	60%
P	mg kg^{-1}	0.00	35.25	28.95	224.50	29.16	83%
K	mg kg^{-1}	0.00	237.96	192.40	2184.60	183.75	77%
CEC	cmol (+) kg^{-1}	0.00	15.77	13.20	83.50	10.11	64%

Notes: SD: standard deviation; CV: coefficient of variation.

The correlation matrix for the ten soil properties is shown in Figure 3. Soil clay content had a strong positive correlation with CEC. The high specific surface area and inherent negative charge of the soil clay content can contribute to cation adsorption [29]. Conversely, soil sand content and silt content presented a weaker correlation with CEC, largely because of their smaller surface areas and diminished charge properties. In addition, a strong correlation was observed between OC and N. This can be attributed to their close interactions with various factors, including biological activity, chemical structure, ecological cycling, and agricultural management practices [30,31].

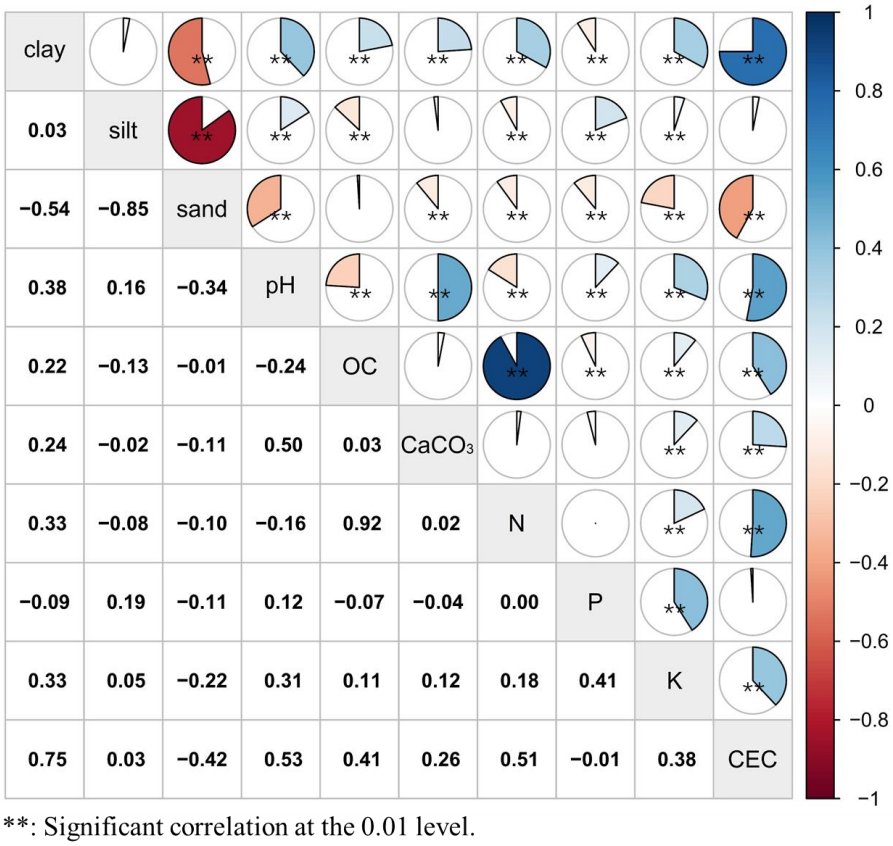

Figure 3. Correlation matrix of the soil properties.

Strong correlations between OC and N can cause collinearity in the SFI estimation process. N is a key nutrient that is essential for plant growth and development. Understanding soil nitrogen content is critically important in agricultural management, especially when precise nitrogen fertilization is required. Therefore, we decided to exclude OC and retain other soil properties when calculating the SFI. This exclusion did not cause the covariance of OC to be lost because this property was strongly correlated with N.

3.2. Development of the Soil Fertility Index

Bartlett's sphericity test was conducted for nine soil properties (clay content, silt content, sand content, pH, $CaCO_3$, N, P, K, and CEC). The factor rotation method was the maximum variance method with the characteristic root threshold set of 1. The results were shown in Figure 4. Four primary components (PC) were extracted, explaining a cumulative variance of 82%. PC1 comprised clay content, N, and CEC; PC2 comprised silt content and sand content; PC3 comprised pH and $CaCO_3$; and PC4 comprised P and K. The final set of properties used to develop the SFI comprised clay content, silt content, sand content, pH, $CaCO_3$, N, P, K, and CEC. Based on our calculations and analyses, the weights of clay content, silt content, sand content, pH, $CaCO_3$, N, P, K, and CEC in SFI were 12%, 10%, 13%, 12%, 8%, 12%, 9%, 11%, and 13%, respectively.

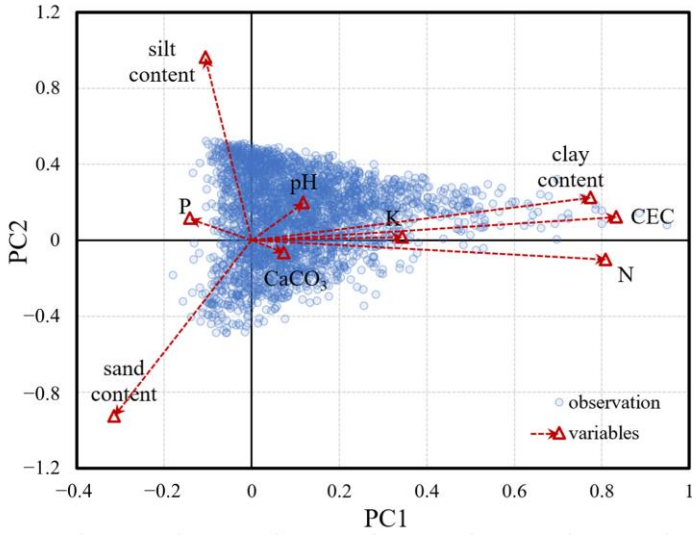

Figure 4. Biplot of factor loadings (PC: principal component).

3.3. Performance of Prediction Models for Soil Properties

Table 2 and Figure 5 present the prediction results for various property models using the validation dataset. SVM models used a radial basis function (RBF) kernel. The RF models contained three key parameters: the number of trees, the minimum number of samples at the terminal nodes, and the number of variables tried at each node, which were set to 500, 5, and 10, respectively. The CNN model primarily consisted of fully connected convolutional and pooling layers. The settings for each layer type were as follows: number of filters: 32; kernel size: 3; pooling window size: 2; and number of neurons: 64. Except for K, for which the CNN model performed the best, the SVM model had the best prediction accuracy for various soil properties among the four models, whereas the RF model showed a relatively poorer performance.

Table 2. Assessment statistics for the different models when predicting the soil properties in the validation dataset.

		Clay Content	Silt Content	Sand Content	pH	$CaCO_3$	N	P	K	CEC
PLSR	R^2	0.72	0.62	0.54	0.82	0.92	0.75	0.32	0.42	0.76
	RMSE	5.28	11.06	13.65	0.43	36.89	0.54	25.43	134.40	4.18
	RPD	1.90	1.60	1.48	2.35	3.45	1.87	1.21	1.30	1.98
	RPIQ	2.65	2.53	2.20	4.21	0.57	2.02	1.45	1.41	2.61
SVM	R^2	0.79	0.72	0.66	0.84	0.95	0.80	0.40	0.48	0.83
	RMSE	4.60	9.30	11.76	0.40	29.12	0.46	24.37	129.34	3.46
	RPD	***2.18***	***1.90***	***1.71***	***2.53***	***4.37***	***2.22***	***1.26***	1.36	***2.40***
	RPIQ	***3.04***	***3.01***	***2.55***	***4.54***	***0.72***	***2.40***	***1.51***	1.47	***3.16***
RF	R^2	0.69	0.60	0.55	0.75	0.82	0.70	0.37	0.41	0.72
	RMSE	5.77	12.45	14.64	0.52	59.59	0.65	24.93	137.52	4.67
	RPD	1.74	1.42	1.38	1.94	2.14	1.56	1.23	1.27	1.78
	RPIQ	2.43	2.25	2.05	3.48	0.35	1.68	1.48	1.38	2.34
CNN	R^2	0.75	0.69	0.66	0.75	0.95	0.68	0.39	0.51	0.75
	RMSE	5.50	11.43	12.51	0.54	38.31	0.62	24.44	125.31	4.47
	RPD	1.82	1.55	1.61	1.87	3.32	1.64	1.26	***1.40***	1.85
	RPIQ	2.54	2.45	2.40	3.35	0.55	1.77	1.51	***1.52***	2.45

Notes: SVM: support vector machine; RF: random forest; CNN: convolutional neural network. The optimal prediction results of each soil property are shown in bold italic font.

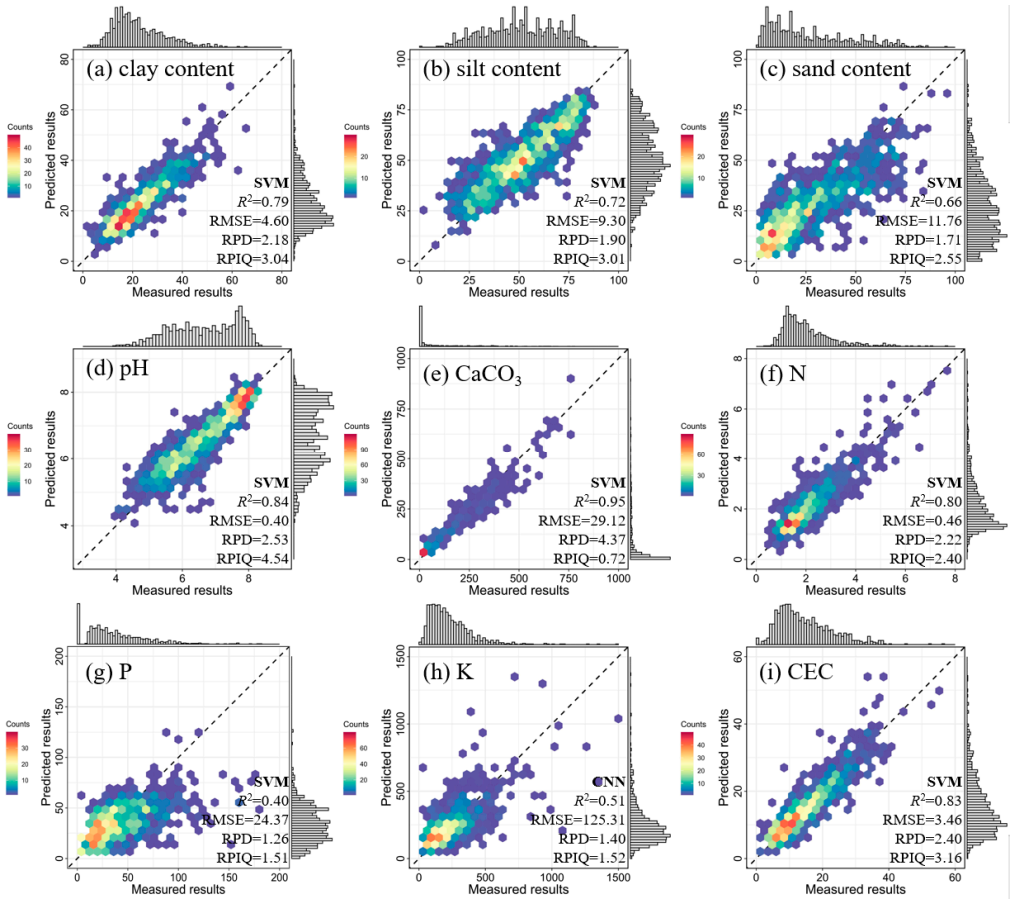

Figure 5. Plots of predicted versus measured values of the different soil properties in the validation dataset using the optimal models. (SVM: support vector machine; CNN: convolutional neural network).

In terms of the RPD, the mean prediction accuracies for the nine soil properties, ranked from highest to lowest, were $CaCO_3$, pH, CEC, clay content, N, silt content, sand content, K, and P. Based on the RPIQ, the rankings were as follows: pH, clay content, CEC, silt content, sand content, N, P, K, and $CaCO_3$. The $CaCO_3$ content was significantly different between the two evaluation indices. The median values for the $CaCO_3$ calibration and validation datasets were both 1.00 g kg^{-1}, and the ranges were 0.00 to 944.00 g kg^{-1} and 0.00 to 770.00 g kg^{-1}, respectively. Both datasets exhibited highly skewed distributions. This severely impacted the RPIQ value, but had less influence on the RPD value. Hence, it can be concluded that the prediction model for $CaCO_3$ had good prediction accuracy (based on the RPD value), but performed poorly when considering the range of data variation (based on the RPIQ value), especially for low $CaCO_3$ values. Overall, clay content, pH, N, and CEC showed the best predictive performances, followed by silt content, sand content, and $CaCO_3$. Although K and P demonstrated the poorest predictive capabilities, they met the basic requirements for the indirect prediction of SFI in subsequent analyses.

3.4. Modeling of SFI Based on VNIR Spectra

Statistical analysis of the measured SFI values revealed a mean value of 0.28, a median value of 0.27, and a CV of 21%. This study used both direct and indirect prediction methods to estimate the SFI values (Table 3). Direct prediction involves constructing a relationship between the spectra and SFI using four different models: PLSR, SVM, RF, and CNN. Among the four models, CNN showed the highest prediction accuracy, with an RPD of 1.63 and an RPIQ of 2.38. In contrast, indirect prediction involves predicting soil properties using spectra and then selecting the best prediction results with which to calculate the SFI

values. Compared to direct prediction, indirect prediction showed a slight improvement in accuracy, with an RPD of 1.74 and an RPIQ of 2.55. The results of both the direct and indirect models demonstrate that the rapid estimation of soil fertility using VNIR is feasible and offers a high level of accuracy.

Table 3. Comparison of the soil fertility index between direct prediction and indirect prediction.

Method		R^2	RMSE	RPD	RPIQ
Indirect model		***0.83***	***0.04***	***1.74***	***2.55***
Direct model	PLSR	0.80	0.05	1.37	2.00
	SVM	0.83	0.05	1.41	2.05
	RF	0.77	0.05	1.41	2.06
	CNN	0.77	0.04	1.63	2.38

Notes: SVM: support vector machine; RF: random forest; CNN: convolutional neural network. The optimal prediction results of SFI are shown in bold italic font.

The number and value ranges for the SFI classification were determined according to the frequency distribution of the comprehensive evaluation result. Soil fertility was classified into five categories: extremely low (SFI < 0.10), low (SFI: 0.10–0.20), medium (SFI: 0.20–0.30), high (SFI: 0.30–0.50), and extremely high (SFI > 0.50). The indirect prediction results for the validation dataset were classified into five classes. Table 4 presents a comparison between the SFI prediction and the measured results for classification. From the perspective of producer accuracy and user accuracy, the classification accuracies for medium and high levels were relatively high, while the performances for the low and extremely high levels were relatively poor. Most of the samples were distributed at the medium and high levels, and only 25 and 2 out of a total of 928 samples were at the low or extremely high levels, resulting in inadequate training data at these levels. The kappa coefficient of 0.63 indicated that the SFI evaluation results of the VNIR and chemical analysis were substantially consistent, suggesting that VNIR could be a reliable method for assessing soil fertility.

Table 4. Comparison of the soil fertility classification between indirect prediction and chemical analysis.

IP \ CA	Extremely Low	Low	Medium	High	Extremely High	Total	User Accuracy
Extremely low	0	0	0	0	0	0	-
Low	0	4	6	0	0	10	40%
Medium	0	21	325	9	0	355	92%
High	0	0	138	418	2	558	75%
Extremely high	0	0	0	5	0	5	0%
Total	0	25	469	432	2	928	-
Producer accuracy	-	16%	69%	97%	0%	-	kappa: 0.63

Notes: IP, indirect prediction; CA, chemical analysis.

4. Discussion

4.1. Capability of Spectroscopy for Soil Properties

In this study, we successfully achieved the prediction of various soil properties using VNIR (consisting the models of PLSR, SVM, RF, and CNN). The values of selectivity ratio (SR) are calculated for individual variables. Figure 6 shows the SR of each soil property. It was considered that the band had more influence on the spectral prediction model when the SR value was higher. Among these properties, clay content, pH, N, CEC, and OC showed the best predictive performance, followed by silt content, sand content, and $CaCO_3$. The prediction of K and P was more challenging, possibly due to their lower concentrations and lack of direct spectral response in the VNIR spectra [32]. Although OC was not applied to the evaluation of SFI, it showed high accuracy in the prediction model based on SVM (

mboxemphR^2 = 0.81, RPD = 2.21, RPIQ = 2.55). The advantages of VNIR in the prediction of OC and N were mainly due to direct spectral responses to the overtones and combinations of N—H, C—H + C—H, and C—H + C—C [9]. Although there is no direct spectral response for pH in the VNIR spectra, its prediction accuracy is high, which may be related to mineral wavelengths [33,34]. The clay content plays a crucial role in soil structure, associated with OH in water and Mg^{-}, Al^{-}, and Fe—OH in mineral lattices [35]. CEC is key to the buffering capacity of soil, which is closely related to clay content and OM [32]. Therefore, VNIR spectroscopy was able to predict the clay content and CEC indirectly [36,37]. The prediction results for various soil properties in this study are basically consistent with previous research [25,38,39].

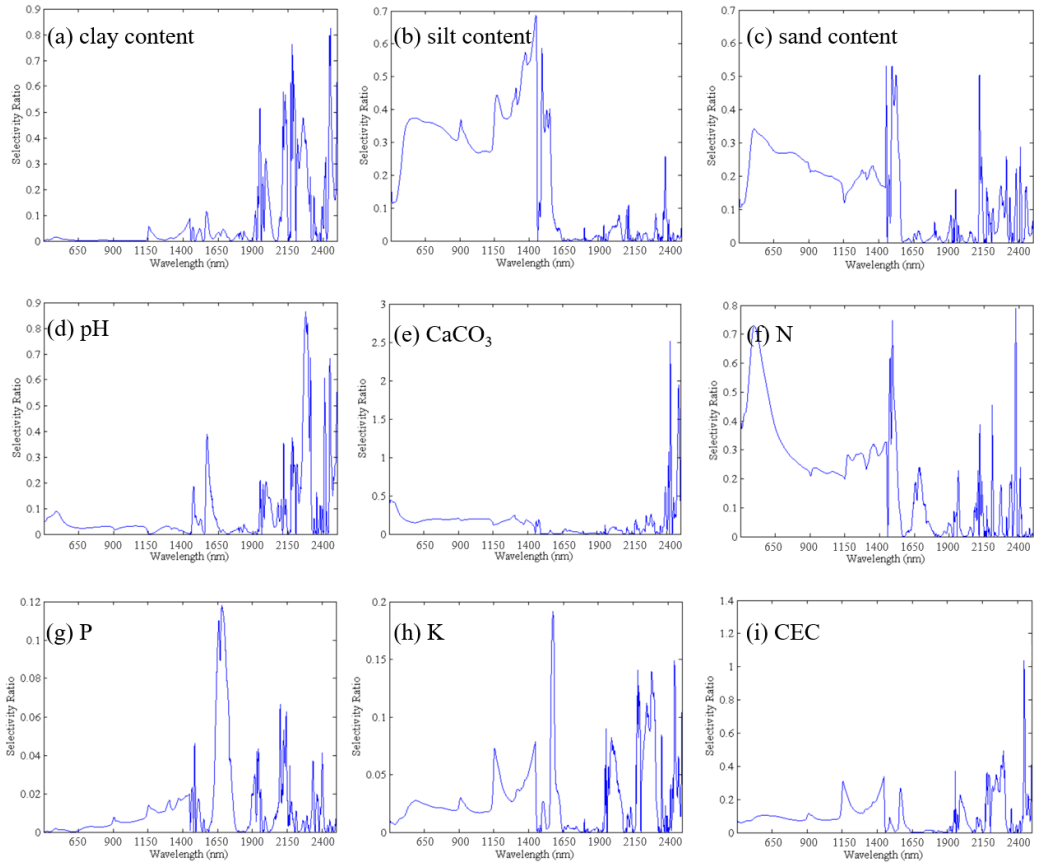

Figure 6. Selectivity ratio of each soil property based on PLSR models.

4.2. Capability of Spectroscopy for SFI Estimation

Due to limitations in the dataset used, SFI developed in this study mainly consisted of soil conventional chemical and physical properties. Although soil biological indicators were not included, it is acknowledged that these indicators significantly influence soil fertility. The determination of weights for the properties was based on statistical analysis, ensuring objectivity in SFI evaluation. However, this might conflict with empirical knowledge and could vary across different regions. Despite these limitations, our study demonstrated the effectiveness of combining VNIR spectroscopy and machine learning models in the assessment of soil fertility. In contrast to other studies, this study did not distinguish soil types and sought a more universally applicable method for assessing soil fertility.

We compared the performance of indirect and direct prediction methods in estimating SFI. The results indicated that both methods achieved high accuracy in SFI estimation, with indirect prediction slightly outperforming direct prediction. This suggests that the accurate prediction of individual soil properties can contribute to the assessment of SFI. The indirect prediction not only had a high SFI prediction accuracy, but also allowed us to obtain specific

soil property values without increasing costs. Current research mainly focuses on using specific soil properties to represent the overall soil fertility, while soil fertility evaluation directly using spectroscopy technology is relatively limited [40,41]. However, the direct prediction method, due to its streamlined and efficient process, shows great potential for the real-time monitoring of regional soil fertility changes. Therefore, whether using direct or indirect prediction, choosing the appropriate method is crucial for devising effective land use strategies based on soil conditions.

4.3. Application of SFI in Precision Agriculture

SFI is a valuable indicator for sustainable productivity in terms of its capability to assess the effectiveness of soil management measures and soil functions [5]. It can be classified into different levels, which can provide fundamental information for precision agriculture management. Therefore, evaluating appropriate thresholds of SFI quickly and effectively is crucial for assessing management measures, which are beneficial for improving crop production while minimizing environmental impacts [42,43].

This study offers rapid, low-cost, and reliable methods for estimating SFI, yet the reliance on point-based measurements limits its broader applicability. To extend these estimations to the regional scale, an integration of point-based data with remote sensing technology is imperative. However, satellite data face challenges in practical application. One of the challenges is the mixed-pixel problem due to mismatch between remote sensing data resolution and soil samples. The constrained spectral bandwidth hampers the capture of soil spectral characteristics as well. These limitations result in SFI estimation with poor accuracy compared with proximal soil sensing methods used in the field [44]. To address these challenges, the fusion of proximal soil sensing with satellite remote sensing data will be attempted in our future research. This integration of data could provide decision makers with more comprehensive and accurate soil surface information, facilitating sustainable land management and utilization.

5. Conclusions

This study provided a comprehensive comparative analysis of soil fertility assessment methods using VNIR spectroscopy. The main conclusions are as follows: (1) VNIR spectra can be effectively employed to predict various soil properties. Clay content, pH, OC, and CEC showed the highest prediction performance, followed by silt content, sand content, N, and CaCO3, whereas K and P had the lowest prediction accuracy. (2) Based on factor analysis, we developed an SFI that integrates nine essential soil properties: clay content, silt content, sand content, pH, $CaCO_3$, N, P, K, and CEC. When comparing direct and indirect prediction models for SFI estimation, the indirect prediction model had a higher accuracy, with RPD = 1.74 and RPIQ = 2.55. (3) The SFI was classified into five categories, with a kappa coefficient of 0.63. This suggests that the SFI evaluation results of the VNIR and the chemical analyses were consistent. Overall, VNIR provides a theoretical foundation for the real-time monitoring of soil fertility changes and the optimization of agricultural practices.

Author Contributions: Formal analysis, X.J.; writing—original draft, X.J.; writing—review and editing, Y.Z., B.H., Y.F. and B.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the National Science Foundation of China (Nos. 42001047, 42201073, 42201054), and the Jiangxi "Double Thousand plan" (No. jxsq202301091).

Data Availability Statement: The data used in this study can be downloaded from the official EUROSTAT website (http://ec.europa.eu/eurostat/web/lucas/data/primary-data/2009, accessed on 8 December 2023).

Conflicts of Interest: The authors have no relevant financial or non-financial interests to disclose.

Abbreviations

SFI: soil fertility index; VNIR: visible-near-infrared spectroscopy; CEC: cation exchange capacity; OM: organic matter; PLSR: partial least squares regression; LUAS: land use and cover area frame survey; OC: organic carbon; $CaCO_3$: carbonates content; N: total nitrogen; P: available phosphorus; K: available potassium; KS: Kennard–Stone; SVM: support vector machine; RF: random forest; CNN: convolutional neural network; PCA: principal component analysis; R^2: coefficient of determination; RMSE: root mean square error; RPD: ratio of prediction to deviation; RPIQ: ratio of performance to interquartile range; *k*: Kappa coefficient; CV: coefficient of variation; PC: primary component; SR: selectivity ratio.

References

1. Abbott, L.K.; Murphy, D.V. *What Is Soil Biological Fertility? Soil Biological Fertility, a Key to Sustainable Land in Agriculture*; Kluwer Academic Publishers: Amsterdam, The Netherlands, 2007.
2. Naumann, M.; Koch, M.; Thiel, H.; Gransee, A.; Pawelzik, E. The importance of nutrient management for potato production part II, plant nutrition and tuber quality. *Potato Res.* **2020**, *63*, 121–137. [CrossRef]
3. Bongiovanni, R.; Lowenberg-Deboer, J. Precision agriculture and sustainability. *Precis. Agric.* **2004**, *5*, 359–387. [CrossRef]
4. Spomer, R.G.; Piest, R.F. Soil productivity and erosion of Iowa loess soils. *Trans. ASAE* **1982**, *25*, 1295–1299. [CrossRef]
5. D'Hose, T.; Cougnon, M.; De Vliegher, A.; Vandecasteele, B.; Viaene, N.; Cornelis, W.; Van Bockstaele, E.; Reheul, D. The positive relationship between soil quality and crop production: A case study on the effect of farm compost application. *Appl. Soil Ecol.* **2014**, *75*, 189–198. [CrossRef]
6. Nocita, M.; Stevens, A.; Wesemael, B.; Aitkenhead, M.; Bachmann, M.; Barthès, B.; Dor, E.B.; Brown, D.J.; Clairotte, M.; Csorba, A.; et al. Soil spectroscopy, An alternative to wet chemistry for soil monitoring. *Adv. Agron.* **2015**, *132*, 139–159.
7. Kuang, B.; Mahmood, H.S.; Quraishi, Z.; Hoogmoed, W.B.; Mouazen, A.M.; Henten, V. Sensing soil properties in the laboratory, in situ, and on-line. *Adv. Agron.* **2012**, *114*, 155–223.
8. Soriano-Disla, J.M.; Janik, L.J.; Viscarra Rossel, R.A.; Macdonald, L.M.; Mclaughlin, M.J. The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. *Appl. Spectrosc. Rev.* **2014**, *49*, 139–186. [CrossRef]
9. Xu, D.Y.; Ma, W.Z.; Chen, S.C.; Jiang, Q.S.; He, K.; Shi, Z. Assessment of important soil properties related to Chinese Soil Taxonomy based on vis–NIR reflectance spectroscopy. *Comput. Electron. Agr.* **2018**, *144*, 1–8. [CrossRef]
10. Munnaf, M.A.; Haesaert, G.; Meirvenne, M.V.; Mouazen, A.M. Site-specific seeding using multi-sensor and data fusion techniques, A review. *Adv. Agron.* **2020**, *161*, 241–323.
11. Viscarra Rossel, R.A.; Rizzo, R.D.; Demattе, J.A.M.; Behrens, T. Spatial modeling of a soil fertility index using visible-near-infrared spectra and terrain properties. *Soil Sci. Soc. Am. J.* **2010**, *74*, 1293–1300. [CrossRef]
12. Askari, M.S.; Cui, J.; O'Rourke, S.M.; Holden, N.M. Evaluation of soil structural quality using VIS-NIR spectra. *Soil Tillage Res.* **2015**, *146*, 108–117. [CrossRef]
13. Yang, M.H.; Abdul, M.; Zhao, X.M.; Guo, X. Assessment of a soil fertility index using visible and near-infrared spectroscopy in the rice paddy region of southern china. *Eur. J. Soil Sci.* **2020**, *71*, 615–626. [CrossRef]
14. Munnaf, M.A.; Mouazen, A.M. Development of a soil fertility index using on-line Vis-NIR spectroscopy. *Comput. Electron. Agr.* **2021**, *188*, 106341. [CrossRef]
15. Gergely, T.; Arwyn, J.; Luca, M.; Christine, A.; Cristiano, B.; Florence, C.; Delphine, B.; Rannveig, G.; Ciro, G.; Tamás, H.; et al. *Lucas Topoil Survey-Methodology, Data and Results*; Publications Office of the European Union: Luxembourg, 2013.
16. Andrews, S.S.; Karlen, D.L.; Mitchell, J.P. A comparison of soil quality indexing methods for vegetable production systems in Northern California. *Agr. Ecosyst. Environ.* **2002**, *90*, 25–45. [CrossRef]
17. Kaiser, H.F. The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **1958**, *23*, 187–200. [CrossRef]
18. Li, S.; Ji, W.J.; Chen, S.C.; Peng, J.; Zhou, Y.; Shi, Z.; Ben-Dor, E.; Kellndorfer, J.; Thenkabail, P.S. Potential of VIS-NIR-SWIR spectroscopy from the chinese soil spectral library for assessment of nitrogen fertilization rates in the paddy-rice region, China. *Remote Sens.* **2015**, *7*, 7029–7043. [CrossRef]
19. Li, S.; Shi, Z.; Chen, S.C.; Ji, W.J.; Zhou, L.Q.; 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]
20. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
21. Zhang, Y.; Yu, Q.; Liu, X.; Suganthan, P.N. A survey on multi-class multi-kernel learning for support vector machines. *Neural Networks* **2021**, *141*, 297–313.
22. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
23. Cutler, D.R.; Edwards, T.C.J.; Beard, K.H.; Cutler, A.; Hess, K.T.; Gibson, J.; Lawler, J.J. Random forests for classification in ecology. *Ecology* **2007**, *88*, 2783–2792. [CrossRef] [PubMed]
24. Lecun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]

25. Shi, Z.; Wang, Q.L.; Peng, J.; Ji, W.J.; Liu, H.J.; 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]
26. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174. [CrossRef]
27. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2014.
28. Hu, B.F.; Chen, S.C.; Hu, J.; Xia, F.; Xu, J.F.; Li, Y.; Shi, Z. Application of portable XRF and VNIR sensors for rapid assessment of soil heavy metal pollution. *PLoS ONE* **2017**, *12*, e0172438. [CrossRef] [PubMed]
29. Qafoku, N.P. Ion exchange reactions in soils, principles and applications. *Adv. Agron.* **2018**, *150*, 1–50.
30. Batjes, N.H. Total carbon and nitrogen in the soils of the world. *Eur. J. Soil Sci.* **1996**, *47*, 151–163. [CrossRef]
31. Davidson, E.A.; Janssens, I.A. Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. *Nature* **2006**, *440*, 165–173. [CrossRef]
32. Stenberg, B.; Viscarra Rossel, R.A.; Mouazen, A.M.; Wetterlind, J. Visible and near infrared spectroscopy in soil science. *Adv. Agron.* **2010**, *107*, 163–215.
33. Viscarra Rossel, R.A.; Behrens, T. Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma* **2010**, *158*, 46–54. [CrossRef]
34. Ji, W.J.; Shi, Z.; Huang, J.Y.; Li, S. In situ measurement of some soil properties in paddy soil using visible and near-infrared spectroscopy. *PLoS ONE* **2014**, *9*, e105708. [CrossRef]
35. Ben-Dor, E.; Banin, A. Near-infrared analysis as a rapid method to simultaneously evaluate several soil properties. *Soil Sci. Soc. Am. J.* **1995**, *59*, 364–372. [CrossRef]
36. Brown, D.J.; Shepherd, K.D.; Walsh, M.G.; Dewayne Mays, M.; Reinsch, T.G. Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma* **2006**, *132*, 273–290. [CrossRef]
37. Viscarra Rossel, R.A.; Behrens, T.; Ben-Dor, E. A global spectral library to characterize the worlds soil. *Earth-Sci. Rev.* **2016**, *155*, 198–230. [CrossRef]
38. Bilgili, A.V.; van Es, H.M.; Akbas, F.; Durak, A.; Hively, W.D. Visible-near infrared reflectance spectroscopy for assessment of soil properties in a semi-arid area of Turkey. *J. Arid. Environ.* **2010**, *74*, 229–238. [CrossRef]
39. Abdi, D.; Tremblay, G.F.; Ziadi, N.; Bélanger, G.; Parent, L. Predicting soil phosphorus-related properties using near-infrared reflectance spectroscopy. *Soil Sci. Soc. Am. J.* **2012**, *76*, 2318. [CrossRef]
40. Idowu, O.; Es, H.; Abawi, G.; Wolfe, D.; Ball, J.; Gugino, B.; Moebius, B.; Schindelbeck, R.; Bilgili, A. Farmer-oriented assessment of soil quality using field, laboratory, and VNIR spectroscopy methods. *Plant Soil* **2008**, *307*, 243–253. [CrossRef]
41. Kinoshita, R.; Moebius-Clune, B.N.; van Es, H.M.; Hively, W.D.; Bilgilis, A.V. Strategies for soil quality assessment using visible and near-infrared reflectance spectroscopy in a western Kenya chronosequence. *Soil Sci. Soc. Am. J.* **2012**, *76*, 1776–1788. [CrossRef]
42. Ball, B.C.; Batey, T.; Munkholm, L.J. Field assessment of soil structural quality-a development of the Peerlkamp test. *Soil Use Manag.* **2007**, *23*, 329–337. [CrossRef]
43. Askari, M.S.; O'Rourke, S.M.; Holden, N.M. Evaluation of soil quality for agricultural production using visible-near-infrared spectroscopy. *Geoderma* **2015**, *243–244*, 80–91. [CrossRef]
44. Ben-Dor, E.; Taylor, R.G.; Hill, J.; Demattê, J.A.M.; Sommer, S. Imaging spectrometry for soil applications. *Adv. Agron.* **2008**, *97*, 321–392.