agronomy

MDPI

Article

Mapping Soil Organic Matter Content Based on Feature Band Selection with ZY1-02D Hyperspectral Satellite Data in the Agricultural Region

Hengliang Guo [1], Rongrong Zhang [2], Wenhao Dai [2], Xiaowen Zhou [2], Dujuan Zhang [1], Yaohuan Yang [1,*] and Jian Cui [3,4,*]

[1] National Supercomputing Center in Zhengzhou, Zhengzhou University, Zhengzhou 450001, China
[2] School of Geoscience and Technology, Zhengzhou University, Zhengzhou 450001, China
[3] Henan Institute of Geological Survey, Zhengzhou 450001, China
[4] National Engineering Laboratory Geological Remote Sensing Center for Remote Sensing Satellite Application, Zhengzhou 450001, China
* Correspondence: yangyaohuan2005@163.com (Y.Y.); cuijianxy@gmail.com (J.C.); Tel.: +86-158-3719-6021 (Y.Y.); +86-186-1371-9585 (J.C.)

Abstract: Soil organic matter (SOM) is an essential nutrient for crop growth and development. Hyperspectral satellite images with comprehensive spectral band coverage and high spectral resolution can be used to estimate and draw a spatial distribution map of SOM content in the region, which can provide a scientific management basis for precision agriculture. This study takes Xinzheng City, Henan Province's agricultural area, as the research object. Based on ZY1-02D hyperspectral satellite image data, the first derivative of reflectance (FDR) was processed on the original reflectance (OR). The SOM characteristic spectral bands were extracted using the correlation coefficient (CC) and least absolute shrinkage and selection operator (Lasso) methods. The prediction model of SOM content was established by multiple linear regression (MLR), partial least squares regression (PLSR), and random forest (RF) algorithms. The results showed that: (1) FDR processing can enhance SOM spectral features and reduce noise; (2) the Lasso feature band extraction method can reduce the model's input variables and raise the estimation precision; (3) the SOM content prediction ability of the RF model was significantly better than that of the MLR and PLSR models. The FDR-Lasso-RF model was the best SOM content prediction model, and the validation set R^2 = 0.921, MAE_V = 0.512 g/kg, $RMSE_V$ = 0.645 g/kg; (4) compared with laboratory hyperspectral data-SOM prediction methods, hyperspectral satellite data can achieve accurate, rapid, and large-scale SOM content prediction and mapping. This study provides an efficient, accurate, and feasible method for predicting and mapping SOM content in an agricultural region.

Keywords: soil organic matter; hyperspectral satellite data; ZY1-02D; spectral characteristic selection; agricultural region

Citation: Guo, H.; Zhang, R.; Dai, W.; Zhou, X.; Zhang, D.; Yang, Y.; Cui, J. Mapping Soil Organic Matter Content Based on Feature Band Selection with ZY1-02D Hyperspectral Satellite Data in the Agricultural Region. *Agronomy* **2022**, *12*, 2111. https://doi.org/10.3390/agronomy12092111

Academic Editors: José Ramón Rodríguez-Pérez and Shawn C. Kefauver

Received: 14 August 2022
Accepted: 3 September 2022
Published: 5 September 2022

1. Introduction

Organic matter in the soil (SOM) is among the crucial nutrients for crop growth [1,2]. How to assess SOM rapidly and effectively is one of the difficult problems faced by the development of modern agriculture [3]. Traditional SOM measurements rely mainly on field collection of soil samples and laboratory chemical analysis [4]. However, the cost of observation is high, time-consuming, and labor-intensive [5,6], and is incapable of providing dynamic and detailed observation data [7,8]. In addition, the soil sampling process can damage the surface [9], adversely affecting agricultural production.

Remote sensing technology provides a timely, reliable, and non-destructive monitoring method for SOM content estimation [10]. Some researchers anticipate SOM content

using non-imaging hyperspectral approaches, which refer to the use of portable spectrometers to obtain soil spectra and estimates in combination with relevant models [11,12]. In previous studies, linear models such as multiple linear regression (MLR) and partial least squares regression (PLSR) [10,13,14], and machine learning models such as random forest (RF) [15–18] are widely used for SOM in estimation research. In real-time, the soil ground hyperspectral model can quickly and efficiently estimate the SOM content [3,19,20]. However, the observation scale is small, and predicting the SOM content over a vast area is impossible. Many academics additionally forecast SOM content using imaging remote sensing technologies, such as Landsat8 and Sentinel-2, and other multispectral satellite data [21–23]. However, multispectral remote sensing images have few bands and low spectral resolution. For example, the spectral coverage range of the Landsat8 image is 430–1251 nm, a total of 11 spectral bands. The spectral coverage of the Sentinel-2 image is 443–2190 nm, with a total of 12 spectral bands. The spectral resolution of multispectral images is low, and since the detailed characteristic information of soil components cannot be highlighted, unsatisfactory results are obtained for SOM estimation [24]. Hyperspectral satellite remote sensing images have the advantages of high spectral resolution and comprehensive spectral band coverage. It offers a path to provide the detailed geographic distribution of SOM content with excellent precision over large areas [25]. For example, the ZY1-02D hyperspectral satellite can simultaneously acquire spectral information in 166 bands ranging from visible light (400 nm) to short-wave infrared (2500 nm). Recently, studies have shown that the ZY1-02D hyperspectral data constructed based on spectral indices have great potential for SOM content prediction [4]. However, few studies on SOM content prediction use ZY1-02D hyperspectral data. Therefore, it is necessary to further explore the SOM content prediction method based on this data.

However, the rich spectral information in hyperspectral imagery may contain a lot of noise [25]. Studies have found that spectral derivative processing can reduce noise information [26,27], highlight and enhance the subtle spectral information of soil components, and improve the accuracy of prediction models [28,29]. In addition, some scholars believe that selecting a strong information band to build a spectral analysis model can remove redundant variables and improve model accuracy and operating efficiency [13,30,31]. The correlation coefficient (CC) method is commonly used for feature band screening. It can obtain the degree of association between each band with the soil's attributes for characteristic extract. The least absolute shrinkage and selection operator (Lasso) is a feature extraction algorithm [32]. Studies have shown that, compared with other variable selection methods, lasso-based variable selection methods can ensure the accuracy of model predictions. Furthermore, the model has fewer input variables and runs faster [33].

A high-precision SOM content prediction model was established to extract the best spectral band as a predictor variable. This study used first derivative processing to highlight SOM characteristic spectral information in ZY1-02D hyperspectral data. Feature spectral extraction was performed using the CC and Lasso methods. Various models were established to forecast the research area's SOM content.

2. Materials

2.1. Study Area

The research location, Xinzheng City (34°16′~34°39′ N, 113°30′~113°54′ E), is positioned in the middle of Henan Province, China (Figure 1), and is one of the prominent grain-producing areas in Henan Province. The average ground elevation is 108 m, with the terrain being higher in the west and lower in the east. It has the characteristics of a warm continental monsoon climate with distinctive seasons. The average yearly temperature and precipitation are 14.3 degrees Celsius and 676.1 mm, respectively. The study area covers 884.592 square kilometers, and more than half (58.59%) of the land is dedicated to agriculture. The main crop types include wheat, corn, and soybeans.

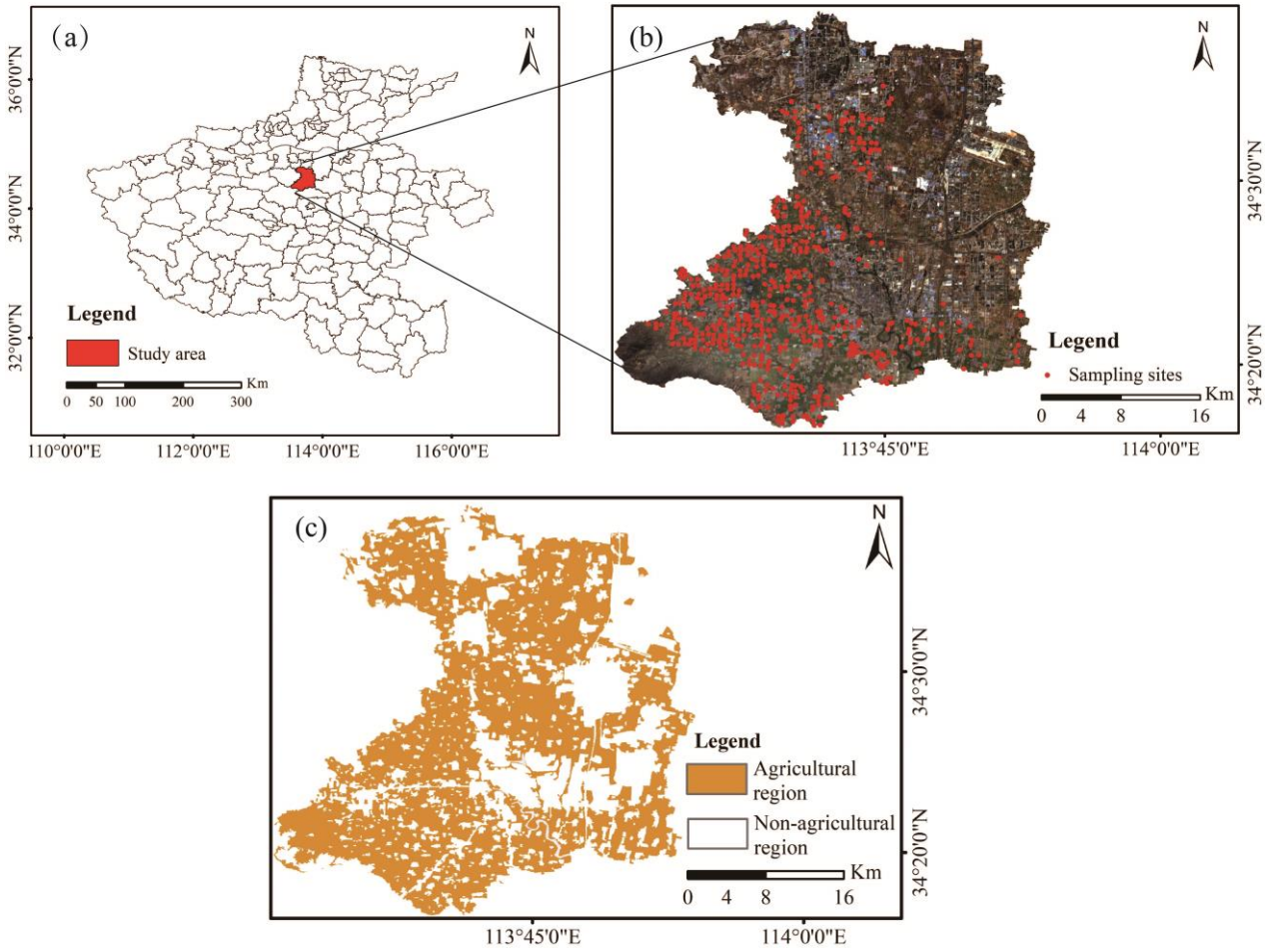

Figure 1. Map of the study area ((**a**): map of Henan Province; (**b**) ZY1-O2D hyperspectral image and location of the study area's sampling locations; (**c**) the study area's distribution of agricultural areas).

2.2. Gathering and Preparation of the Soil Samples

In March 2021, in the study region, a total of 539 soil samples were obtained. Five soil samples were collected in a 30 × 30 m square using the 5-point sampling method for each sampling point, and the collection depth was 0–20 cm. The soil of each sub-sample collected was broken into pieces, and the roots, straws, stones, and other debris were picked out and mixed thoroughly. After that, 1.0–1.5 kg of the soil samples was collected by the quartering method. The samples were naturally air-dried in the room and passed through a 1.988 mm sieve after drying. The total weight of the under-sieve portion was more than 300 g. Utilizing the concentrated sulfuric acid–potassium dichromate volumetric heating technique, the SOM content of all soil samples was measured [34].

2.3. Hyperspectral Satellite Data Acquisition, Pre-Processing, and First Derivative Processing

The Advanced Hyperspectral Imager (AHSI) of ZY1-02D obtained the hyperspectral satellite image that was used in this investigation (Figure 1b) on 28 January 2021. Table 1 contains the specific information about the AHSI hyperspectral sensor. The AHSI has a spatial resolution of 30 m, a stripe breadth of 60 km, a 400–2500 nm spectral range, and a total of 166 spectral channels. The visible and near-infrared (VNIR) spectral range is 76 bands, and the short-wave infrared (SWIR) spectral range is 90 bands. The spectral resolutions of VNIR and SWIR are 10 nm and 20 nm, and three overlapping bands (77–79) were removed in this study.

Table 1. ZY1-02D AHSI hyperspectral sensor specifications.

Satellite Payloads	ZY-1-02D
Launch Time	12 December 2019
Number of Bands	76 (VNIR), 90 (SWIR)
Spectral Range (nm)	400–2500
Spectral Resolution (nm)	10 (VNIR), 20 (SWIR)
Spatial Resolution (m)	30
Swath Width (km)	60
Revisit Cycle (d)	3

This study used ENVI5.3 (Environment for Visualizing Images 5.3, Harris Geospatial Corporation, Broomfield, CO, USA) for orthorectification, radiometric correction, and atmospheric correction. Savitzky–Golay (SG) filtering was used to remove background noise information and smooth the spectrum [35]. After SG filtering and smoothing, the reflectance data were used as the original reflectance (OR) data of soil samples for spectral processing and predictive modeling. The spectral reflectance SG smoothing was performed in MATLAB R2018a (MathWorks Inc., Natick, MA, USA).

The first derivative of reflectance (FDR) processing can highlight the reflection and absorption characteristics of soil components in the spectral curve and enhance the characteristic spectral bands of SOM [36]. It also eliminates the effects of background noise in irrelevant bands, thereby reducing systematic errors caused by instrumentation [37]. FDR processing was implemented in MATLAB R2018a (MathWorks Inc., Natick, MA, USA).

3. Methods

Figure 2 displays the study's flow chart. It consists of three main steps: (1) collect soil samples for chemical analysis; (2) preprocess ZY1-02D hyperspectral images; and (3) SOM characteristic spectrum selection, modeling, and mapping.

3.1. Characteristic Spectral Band Selection

3.1.1. Correlation Coefficient (CC)

This method refers to the correlation analysis based on the Pearson coefficient between the reflectivity of each band and the soil properties to determine the soil parameters' correlation with each band as well as the correlation coefficient curve. CC is currently the most widely used method for screening soil attribute characteristic bands. In this study, the correlation coefficient method's standard for choosing feature bands was level of significance $p < 0.01$, and the correlation analysis operation was completed in SPSS26 (Statistical Product and Service Solutions). The following formula can be used to determine the Pearson coefficient:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \quad (1)$$

where r represents the Pearson correlation coefficient, n represents the number of samples, x_i and y_i represent the spectral reflectance and soil properties of each band, respectively, $\overline{x}$ and $\overline{y}$ indicate the mean value of spectral reflectance and soil properties in each band.

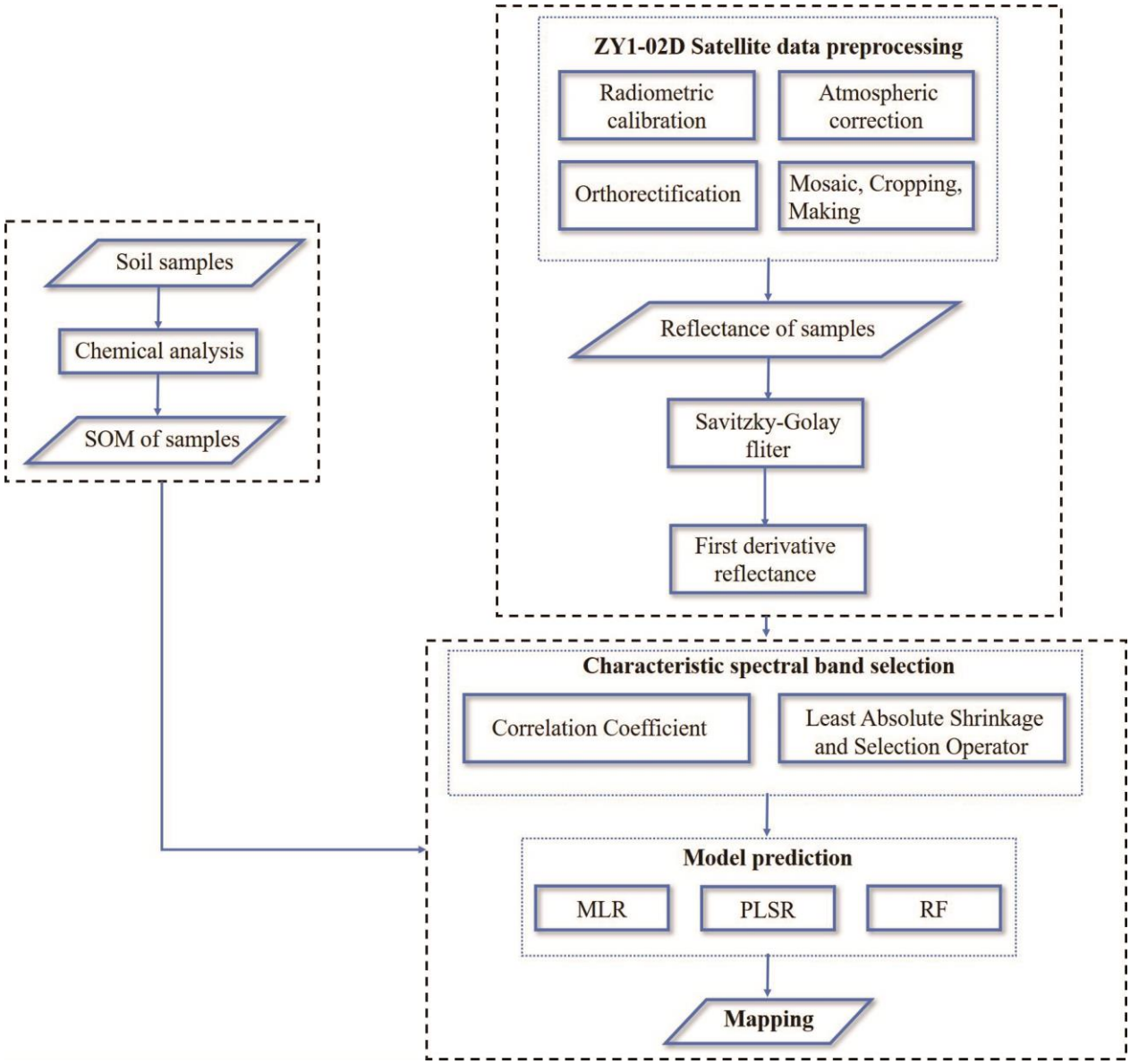

Figure 2. Flow chart.

3.1.2. Least Absolute Shrinkage and Selection Operator (Lasso)

The Lasso algorithm is a feature selection method based on a linear regression model for feature selection by selecting and compressing variables, which can effectively prevent the problem of overfitting [33]. The basic idea of the algorithm is that when the total of the absolute values of the regression coefficients is below a predetermined threshold, it minimizes the residual sum of squares. To decrease the size of the feature space, the coefficients of variables with low correlation are compressed to 0, and then these feature variables are removed. Equation (2) is the mathematical expression of the minimum residual sum of squares of the Lasso feature selection method. Lasso feature selection is performed in Python 3.8.

$$\underset{\rho}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} x_{ij} \beta_j \right)^2 \right\} \text{ subject to } \sum_{j=1}^{m} \left| \beta_j \right| \leq \gamma \quad (2)$$

where β_j is the regression coefficient, n is the number of samples, m is the number of features, x_{ij} is the spectral band reflectance, y_i is the SOM content of each sampling point, γ is the threshold.

3.2. Prediction Models

3.2.1. Multiple Linear Regression (MLR)

MLR, which uses regression equations to measure the linear relationship between the dependent variable and two or more independent variables, is a common SOM estimate algorithm. The fundamental goal is to identify the mathematical formula that most accurately captures the connection between the independent and dependent variables [38].

3.2.2. Partial Least Squares Regression (PLSR)

PLSR is a mathematical optimization technique that, by reducing the sum of squared errors, determines the optimal function match for a piece of data and, in the simplest possible manner, determines some truth values that are utterly unknown. It can prevent overfitting, improve the explanatory power of dependent variables, and more effectively choose the input variables with the best explanatory power [39,40]. Therefore, PLSR is widely used in SOM estimation [41,42].

3.2.3. Random Forest (RF)

The RF model is a decision tree-based ensemble regression technique. Each split node in a decision tree is chosen at random from n inputs to completely partition the variable space [25]. RF is a nonlinear machine learning model, which belongs to a major branch of machine learning. It is insensitive to the multicollinearity of variables and has a wide range of applications in nonlinear problems. Studies have shown that the RF prediction model is superior to other machine learning algorithms [43]. This study establishes the number of decision trees (ntree) and the number of split nodes following experimental testing (mtry), two critical parameters of the RF model, to ntree = 400 and mtry = 2. All the predictive models were built in Python 3.8.

3.3. Model Accuracy Evaluation

In this work, the dataset is split using the sklearn.model_selection import train_test_split module in Python 3.8. In order to ensure the consistency of data input of different estimation models, this study set the random state parameter to 0. 539 samples were divided at random into calibration and validation sets in a ratio of 4:1. The dataset was used as fixed data input for different model combinations to select the best SOM estimation model combination. The coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE) were used as metrics to assess how well various models performed [10]. The related equation is this:

$$R^2 = 1 - \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (4)$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (5)$$

where n is the number of samples, y_i is the SOM observed value of sample i, $\hat{y}_i$ is the SOM predicted value of sample i, the greater the R^2 value, the lower the RMSE and MAE, and the more accurate the model.

3.4. Model Stability Evaluation

The d-factor was employed in this study to assess the model's stability. The stability of the model is evaluated using the degree of the d-factor that is near to 0. That is, the model's

stability is greater the smaller the d-factor. The d-factor has a fairly optimal value within the parameter uncertainty range. The following formula is used to compute the d-factor:

$$\overline{d_r} = \frac{1}{n}\sum_{i=1}^{n}(X_{Ui} - X_{Li}) \quad (6)$$

$$d - factor = \frac{\overline{d_r}}{\sigma_X} \quad (7)$$

where $\overline{d_r}$ is the mean of the upper confidence limit X_{Ui} and the lower confidence limit X_{Li}. The number of samples is n, and the measurement's standard deviation is σ_X.

4. Results

4.1. Statistical Description of SOM Content

Table 2 shows the measured value of SOM content and its data analysis results. The lowest value of SOM content of the total soil samples was 12.413 g/kg, the highest value was 28.446 g/kg, and the average value was 20.316 g/kg. The standard deviation (SD) of the overall sample was 2.501 g/kg, and the coefficient of variation (CV) was 12.309%. SD and CV reflect the spatial variability of SOM. The general description of the sample variable is similar for the calibration and validation set samples, indicating that the data set is divided reasonably.

Table 2. Statistical description of SOM content.

Set	N	Max (g/kg)	Min (g/kg)	Mean (g/kg)	SD (g/kg)	CV (%)
Whole set	539	28.446	12.413	20.316	2.501	12.309
Calibration set	431	28.446	12.413	20.195	2.527	12.513
Validation set	108	27.412	13.792	20.797	2.344	11.271

4.2. Spectral Reflectance Characteristics of Soil Samples

4.2.1. Original Reflectance (OR)

Figure 3a shows that, with the exception of samples with an SOM content of 20.688 g/kg, soil spectral reflectance was inversely proportional to SOM content, and spectral shapes were similar. The wavelength ranges from 619 to 877 nm, and as the wavelength increases, reflectance rises swiftly with a peak at 877 nm. The reflectance between 920 and 1106 nm increased slowly with the increase in wavelength and reached its maximum at 1106 nm. Reflectance decreased rapidly with a wavelength increase between 1123–1459 nm, and reflectance decreased with a wavelength increase between 1510 and 2500 nm.

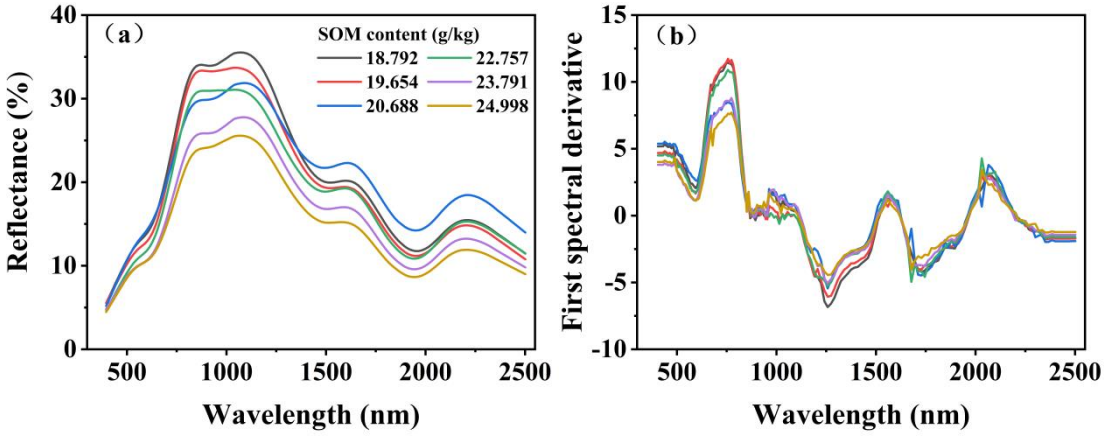

Figure 3. Soil reflectance curves with different soil organic matter content ((**a**) is OR; (**b**) is FDR).

4.2.2. First Derivative Reflectance (FDR)

Studies have shown that FDR can distinguish absorption peaks and enhance weak absorption features [44]. Soil samples with varying SOM content exhibited prominent absorption characteristics as shown in Figure 3b FDR curves at 560–611 nm, 868–912 nm, 1123–1459 nm, and 1695–1812 nm wavelengths. At the same time, FDR also showed prominent reflection peaks between the wavelengths of 687–791 nm, 1543–1593 nm, and 2031–2165 nm. Compared with OR, FDR can highlight more spectral features.

4.3. Characteristic Spectral Band Selection

4.3.1. CC Characteristic Band Choice

A correlation analysis was performed between SOM content and OR and FDR, and the significance test (two-sided) of the correlation coefficient at the level of $p = 0.01$ was performed, as shown in Figure 4 and Table 3. The 133 significantly correlated bands of OR were mainly distributed between 1308–2500 nm wavelengths. There were 149 significant correlation bands in FDR, evenly distributed at wavelengths of 400–2500 nm. The maximum correlation band between OR and SOM content appeared at 1661 nm (correlation coefficient $R = -0.495$). The maximum correlation band between FDR and SOM content appeared at 611 nm (correlation coefficient $R = -0.495$). This study selected the relevant bands with OR and FDR = 0.01 significance levels as characteristic bands.

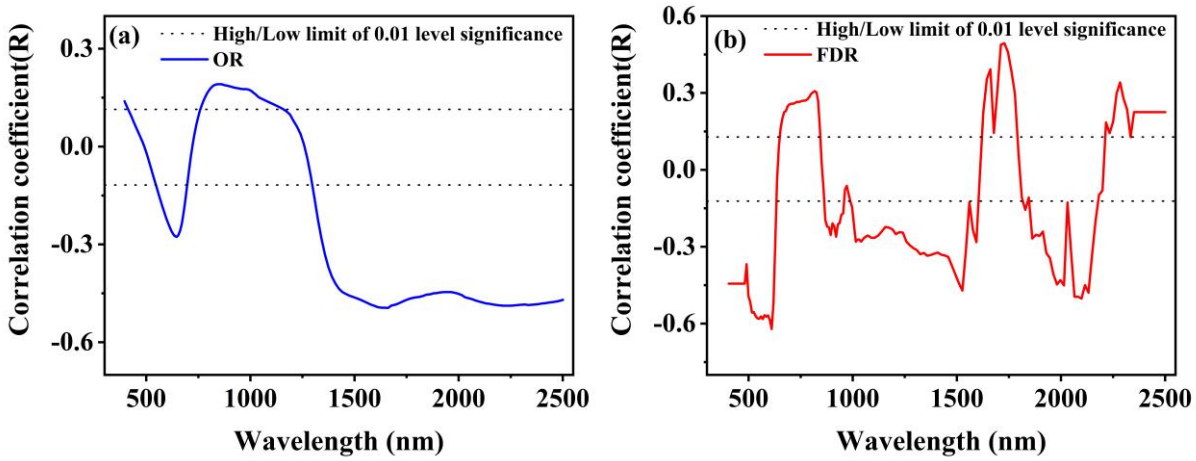

Figure 4. Correlation coefficient of SOM content with spectral wavelength, ((**a**) OR; (**b**) FDR).

Table 3. Correlation coefficient significant correlation band information.

Spectral	Number of Bands	Wavelength Range (nm)	Maximum Correlation Band (nm)	Correlation Coefficient R
OR	133	400–413, 551–697 766–1157, 1308–2500	1661	−0.495
FDR	149	405–628, 654–843 868–954, 988–1594 1627–1779, 1829 1862–2166, 2216–2500	611	−0.621

4.3.2. Lasso Characteristic Spectral Band Choice

Lasso feature selection removed bands with zero β coefficients and retained bands with non-zero β coefficients as feature variables. Figure 5 and Table 4 show that OR-Lasso chose 27 characteristic bands that were uniformly distributed in the 400–2500 nm wavelength range; FDR-Lasso chose 6 characteristic bands that are concentrated in the 500–750 nm and 1750–2100 nm spectral regions.

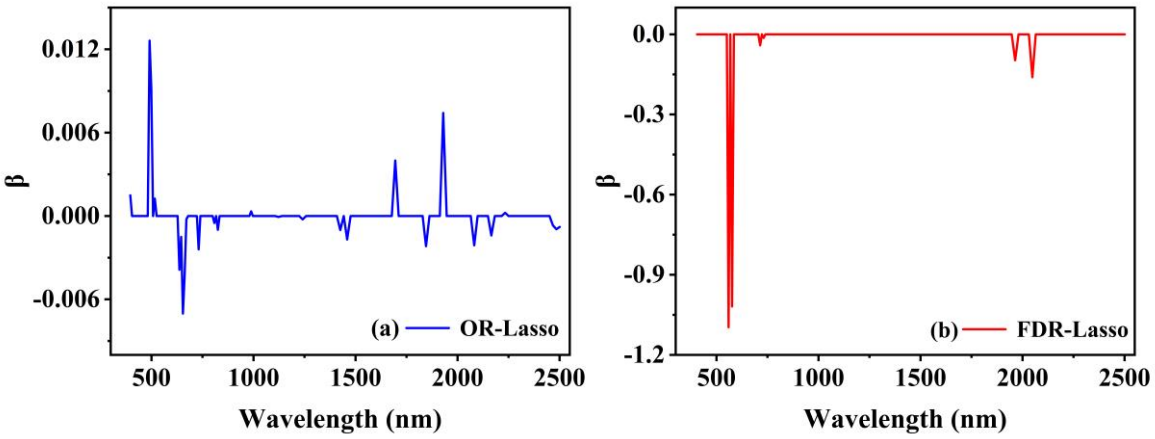

Figure 5. Lasso characteristic spectral band selection ((**a**) OR-Lasso characteristic selection; (**b**) FDR-Lasso characteristic selection).

Table 4. Lasso feature selection band number and wavelength.

Method	Number of Bands	Wavelength (nm)
OR-Lasso	27	400, 490–500, 516, 636–671, 722–731, 808, 825, 988, 1123 1241, 1425, 1459, 1695, 1846, 1930, 2082, 2166, 2233 2468–2500
FDR-Lasso	6	559, 567, 714, 731, 1964, 2048

4.4. The Results of Prediction Models

Table 5 presents the model's projected outcomes for SOM content. From the validation set model, compared to the entire band model and the CC feature selection model, the Lasso feature selection spectral model exhibited greater prediction accuracy. The full-band model and the CC feature selection model differ less in terms of prediction accuracy. The six groups of RF spectral models all had calibration set and validation set R^2 values higher than 0.85, demonstrating the RF models' potent predictive power. FDR-Lasso-RF was the best prediction model. In its validation set, R^2 = 0.921, MAE = 0.512 g/kg, RMSE = 0.645 g/kg. Compared with other models, R^2 was the highest, and MAE and RMSE were the lowest, indicating that the FDR-Lasso-RF model had the optimal SOM prediction performance. In addition to the FDR-Lasso-RF model having a good prediction effect, the OR-Lasso-PLSR and FDR-Lasso-PLSR models also had certain prediction abilities, and their R^2 were 0.448 and 0.411. The MLR model's calibration set and validation set accuracy varied greatly, and its capacity for model estimation was subpar. The RF model's calibration set and validation set both had identical prediction accuracy and a high level of SOM estimation. The prediction outcomes based on FDR were also superior to OR for the RF model.

Figures 6–8 are the fitting diagram of the SOM predicted value and the measured value of the RF model input in different bands. Various RF models had excellent predictive ability. The fitting effect of predicted and measured values based on the Lasso feature selection was better than that of the whole band and CC feature selection. The predicted and measured values of the FDR-Lasso-RF model fit best (Figure 8c,d).

Table 5. The prediction model's outcome statistics.

Model	Method	Calibration Set			Validation Set		
		R_C^2	MAE_C (g/kg)	$RMSE_C$ (g/kg)	R_V^2	MAE_V (g/kg)	$RMSE_V$ (g/kg)
MLR	OR	0.646	1.166	1.503	0.256	1.751	2.203
	OR-CC	0.616	1.203	1.564	0.239	1.712	2.159
	OR-Lasso	**0.453**	**1.471**	**1.863**	**0.296**	**1.536**	**1.925**
	FDR	0.625	1.202	1.545	0.200	1.789	2.279
	FDR-CC	0.601	1.242	1.594	0.172	1.821	2.315
	FDR-Lasso	**0.414**	**1.508**	**1.929**	**0.385**	**1.409**	**1.799**
PLSR	OR	0.330	1.598	2.066	0.307	1.528	1.942
	OR-CC	0.312	1.618	2.094	0.298	1.528	1.954
	OR-Lasso	**0.316**	**1.611**	**2.083**	**0.448**	**1.337**	**1.704**
	FDR	0.316	1.599	2.087	0.286	1.540	1.972
	FDR-CC	0.317	1.599	2.087	0.285	1.539	1.972
	FDR-Lasso	**0.393**	**1.525**	**1.963**	**0.411**	**1.369**	**1.761**
RF	OR	0.900	0.603	0.797	0.880	0.633	0.807
	OR-CC	0.898	0.614	0.806	0.871	0.657	0.838
	OR-Lasso	**0.908**	**0.601**	**0.764**	**0.901**	**0.537**	**0.721**
	FDR	0.913	0.567	0.743	0.897	0.582	0.750
	FDR-CC	0.913	0.570	0.734	0.893	0.590	0.762
	FDR-Lasso	**0.923**	**0.553**	**0.701**	**0.921**	**0.512**	**0.645**

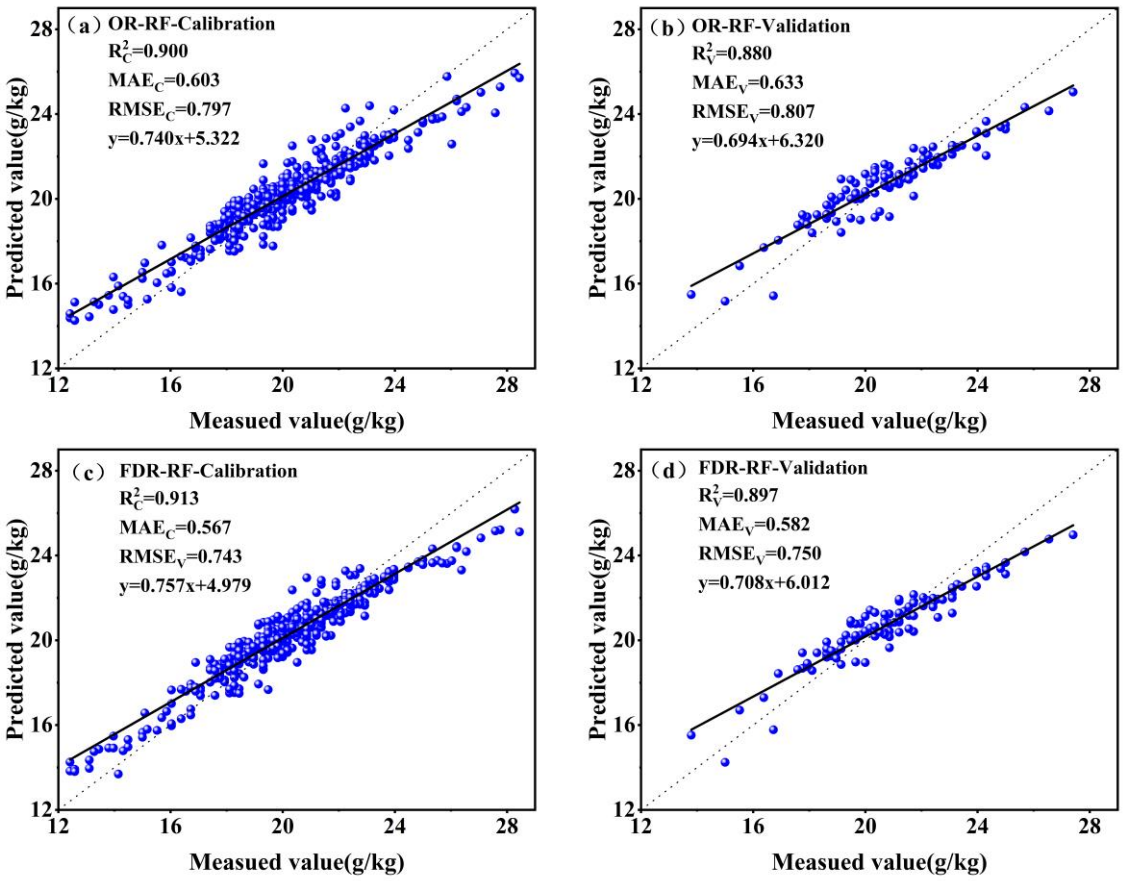

Figure 6. Scatter plot of predicted and measured SOM values for the full-band RF model ((**a**) OR-RF-Calibration; (**b**) OR-RF-Validation; (**c**) FDR-RF-Calibration; (**d**) FDR-RF-Validation).

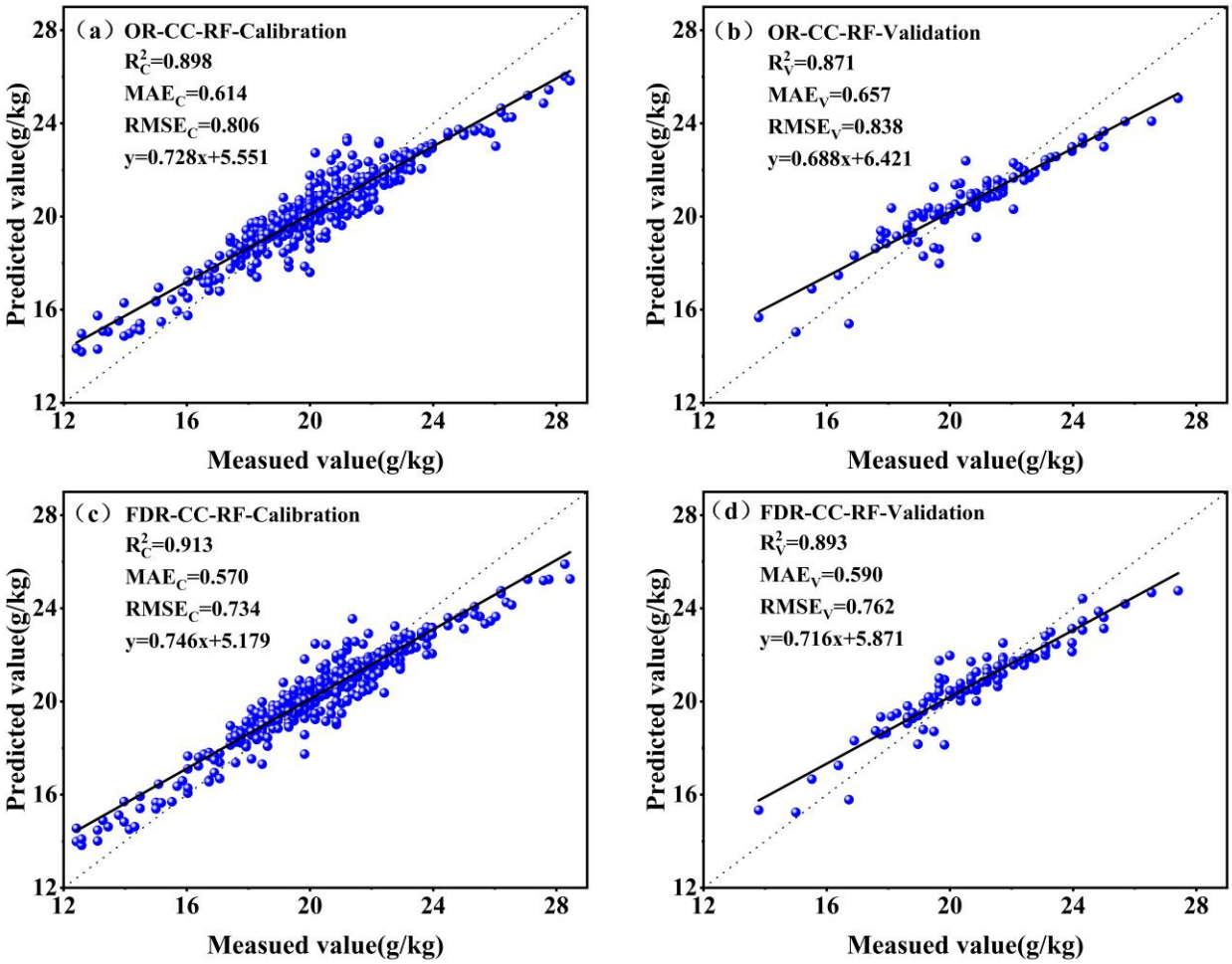

Figure 7. Scatter plot of predicted and measured SOM values for the CC signature band RF model ((**a**) OR-CC-RF-Calibration; (**b**) OR-CC-RF-Validation; (**c**) FDR-CC-RF-Calibration; (**d**) FDR-CC-RF-Validation).

In general, no matter which model was used, the model's prediction accuracy was improved when the Lasso feature selected the band as the model input variable. The model selected based on the CC feature band had the worst accuracy and was not as good as the prediction ability of the whole band. The predictive power of FDR was better than OR for different spectral forms. From the model point of view, RF estimation ability was the best, followed by PLSR, and MLR was the worst. As shown in Figure 8c,d, the FDR-Lasso-RF model still has good prediction accuracy and fitting effect at the extreme value.

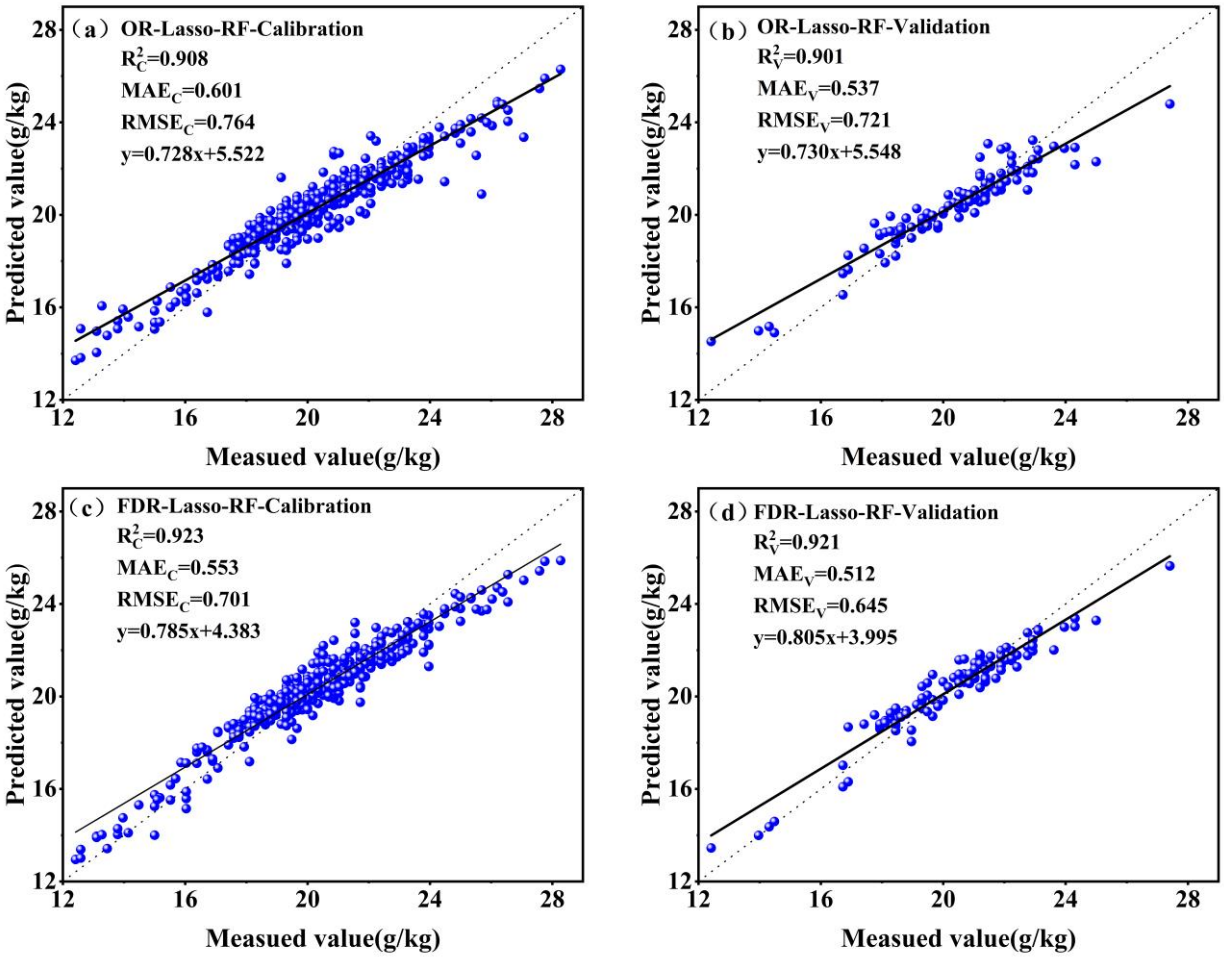

Figure 8. Scatter plot of predicted and measured SOM values for the Lasso signature band RF model ((**a**) OR-Lasso-RF-Calibration; (**b**) OR-Lasso-RF-Validation; (**c**) FDR-Lasso-RF-Calibration; (**d**) FDR-Lasso-RF-Validation).

5. Discussion

According to this study, of the three model algorithms, the RF model had the highest SOM prediction accuracy. Therefore, in the discussion section, the following discussion is made only for predicting SOM content based on the RF model.

5.1. Benefits of FDR Processing

The results (Table 5) show that the predictive power of FDR was better than OR for various RF models. Relevant research has demonstrated that the FDR transform can improve the effective signal in spectral reflectance while reducing the effect of noise [45–47]. This study found that, compared with OR, FDR can highlight the characteristic bands of SOM. The maximum correlation band of FDR and SOM was R = −0.621, and the maximum correlation band of OR was R = −0.495. Together, FDR and the Lasso feature selection technique can minimize the data dimension and boost model accuracy.

5.2. Benefits of Lasso Feature Band Selection

Some scholars have proposed extracting the SOM strong information band for model estimation can reduce the interference of irrelevant variables and improve the prediction accuracy [48,49]. The findings of this study demonstrate that (Table 5) the precision of various prediction models chosen by Lasso bands has increased. The best validation set model accuracy based on Lasso feature selection was R^2 = 0.921, MAE_V = 0.512 g/kg, $RMSE_V$ = 0.645 g/kg. This was because the SOM-independent spectral information con-

tained in the full-band and CC feature selection bands interfered with the estimation results. The Lasso variable selection method can retain feature information, remove noise information, and improve model prediction accuracy.

5.3. Importance Analysis of RF Model Input Variables

The FDR model's capacity for SOM estimation was superior to OR. Consequently, in this research, the important analysis of the input variables of the RF model was carried out based on the FDR spectral data. The importance of RF model variables for different wavelength spectral bands is shown in Figure 9. The bands with the highest importance of RF model variables were concentrated in the wavelength intervals 500–750 nm and 1750–2100 nm. In theory, the greater the importance of the variable, the more outstanding the contribution to the model. The six SOM characteristic bands of FDR selected by the Lasso method in this study were all between the wavelengths of 500–750 nm and 1750–2100 nm. The wavelength range of the characteristic spectrum of SOM in previous studies is consistent [10,13,25,50,51], which is due to the corresponding transformation of some ground spectral reflection bands with the difference of soil organic matter content. Therefore, we can use characteristic spectral bands to estimate the SOM content. The model results show (Table 5) that the FDR-Lasso-RF model had the best SOM content prediction effect. It shows that the RF model has a strong sensitivity to the characteristic spectral bands of SOM. Using the characteristic band as an input variable in the RF model can minimize model complexity while improving forecast accuracy.

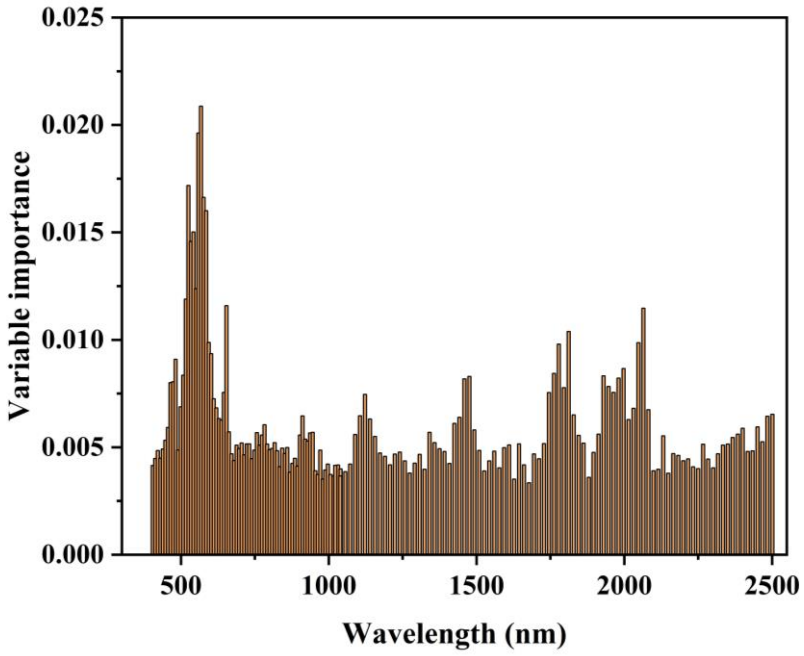

Figure 9. Model predicted contribution importance for RF input variables.

5.4. Optimal Model SOM Mapping

Table 6 displays the findings of the RF model's stability study. The d-factors of the RF models in the six groups of validation sets were all around 0.3, and there was no significant difference. It showed that the FDR-Lasso-RF model still had strong model stability while reducing the model input variables and improving its accuracy.

Table 6. Results of the RF model stability index calculation.

Model	Method	Calibration Set	Validation Set
RF	OR	0.160	0.302
	OR-CC	0.157	0.291
	OR-Lasso	0.156	0.309
	FDR	0.162	0.305
	FDR-CC	0.160	0.296
	FDR-Lasso	0.167	0.307

Note: The smaller the value, the higher the model stability.

The FDR-Lasso-RF model has strong predictive power without increasing model uncertainty (Table 6, Figure 8c,d). Thence, in this study, the best model FDR-Lasso-RF was used for the estimation and SOM content mapping in the research area (Figure 10). The highest and lowest values of SOM content in the study area were 25.012 and 13.331 g/kg. SOM content is higher in the southwest and lower in the center and north. The main reason is that the southwestern region is the Huanghuaihai Plain, with rich soil nutrients and high humus content, and organic matter in the soil can be converted and accumulated efficiently. The central and northern areas are mostly hilly terrain; the soil is poor, humus's fixation and transformation ability are weak, and soil organic matter content is low.

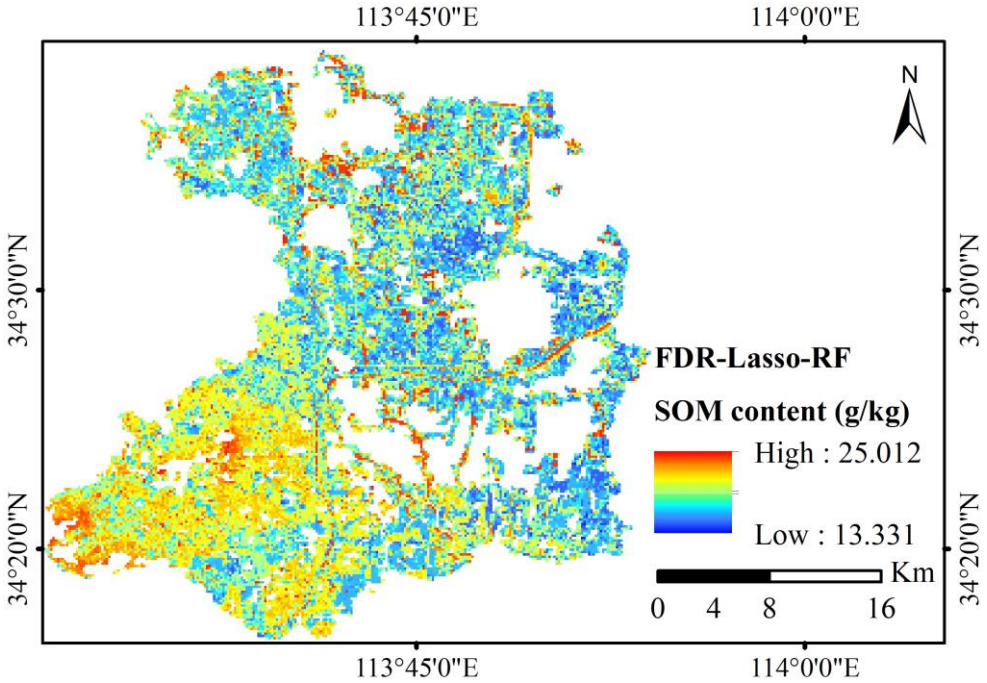

Figure 10. Optimal model predicts SOM content for study area.

6. Conclusions

This study used ZY1−02D hyperspectral satellite data to predict SOM content in agricultural areas. The hyperspectral band reflectance was processed using the first derivative. CC and Lasso were used for SOM spectral information screening. The SOM was estimated using the modeling methods of MLR, PLSR, and RF, and the best model was selected for mapping. The main results are as follows:

(1) In comparison to OR, FDR treatment can increase model fitting accuracy by emphasizing SOM spectral characteristic information in the soil spectrum.
(2) The Lasso feature selection method can effectively extract the SOM feature and spectral bands, reduce the data dimension, highlight key information, and enhance model estimation capabilities.

(3) The RF model provides excellent SOM content prediction and model stability. The FDR-Lasso-RF model is the best prediction model, with $R^2 = 0.921$, $MAE_V = 0.512$ g/kg, and $RMSE_V = 0.645$ g/kg in its validation set.
(4) Hyperspectral satellite data make up for the lack of large-scale observation of laboratory hyperspectral data and provide a feasible method for rapid, large-scale, and accurate prediction and mapping of SOM content in agricultural areas.

Author Contributions: Writing and revision of the manuscript, H.G. and R.Z.; processing and analysis of experimental results, W.D. and X.Z.; data management, D.Z.; experimental data collection and preprocessing, Y.Y.; provision of project funding support, J.C. All authors have read and agreed to the published version of the manuscript.

Funding: The following funds contributed to this study: Major Science and Technology Project of Henan Province (201400210100); Major Science and Technology Project of Henan Province (201400210700); and 2019 Henan Province Natural Science and Technology Project (Henan Natural Letter [2019] No. 373-11).

Data Availability Statement: The data described in this study are accessible from the corresponding author upon request. The data are not publicly available due to privacy.

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

References

1. Zhao, Y.; Xu, X.; Tian, K.; Huang, B.; Hai, N. Comparison of sampling schemes for the spatial prediction of soil organic matter in a typical black soil region in China. *Environ. Earth Sci.* **2016**, *75*, 4. [CrossRef]
2. Tian, J.; Lu, S.; Fan, M.; Li, X.; Kuzyakov, Y. Labile soil organic matter fractions as influenced by non-flooded mulching cultivation and cropping season in rice–wheat rotation. *Eur. J. Soil Biol.* **2013**, *56*, 19–25. [CrossRef]
3. Liu, J.; Dong, Z.; Xia, J.; Wang, H.; Xie, J. Estimation of soil organic matter content based on CARS algorithm coupled with random forest. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2021**, *258*, 119823. [CrossRef] [PubMed]
4. Yang, Y.; Shang, K.; Xiao, C.; Wang, C.; Tang, H. Spectral Index for Mapping Topsoil Organic Matter Content Based on ZY1-02D Satellite Hyperspectral Data in Jiangsu Province, China. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 111. [CrossRef]
5. 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]
6. Wei, L.; Yuan, Z.; Wang, Z.; Zhao, L.; Zhang, Y.; Lu, X.; Cao, L. Hyperspectral Inversion of Soil Organic Matter Content Based on a Combined Spectral Index Model. *Sensors* **2020**, *20*, 2777. [CrossRef]
7. Gruba, P.; Socha, J.; Błońska, E.; Lasota, J. Effect of variable soil texture, metal saturation of soil organic matter (SOM) and tree species composition on spatial distribution of SOM in forest soils in Poland. *Sci. Total Environ.* **2015**, *521–522*, 90–100. [CrossRef]
8. Caddeo, A.; Marras, S.; Sallustio, L.; Spano, D.; Sirca, C. Soil organic carbon in Italian forests and agroecosystems: Estimating current stock and future changes with a spatial modelling approach. *Agric. For. Meteorol.* **2019**, *278*, 107654. [CrossRef]
9. Shi, T.; Liu, H.; Chen, Y.; Wang, J.; Wu, G. Estimation of arsenic in agricultural soils using hyperspectral vegetation indices of rice. *J. Hazard. Mater.* **2016**, *308*, 243–252. [CrossRef]
10. Yu, Q.; Yao, T.; Lu, H.; Feng, W.; Xue, Y.; Liu, B. improving estimation of soil organic matter content by combining Landsat 8 OLI images and environmental data: A case study in the river valley of the southern Qinghai-Tibet Plateau. *Comput. Electron. Agric.* **2021**, *185*, 106144. [CrossRef]
11. Nawar, S.; Mouazen, A.M. Optimal sample selection for measurement of soil organic carbon using on-line vis-NIR spectroscopy. *Comput. Electron. Agric.* **2018**, *151*, 469–477. [CrossRef]
12. Allory, V.; Cambou, A.; Moulin, P.; Schwartz, C.; Cannavo, P.; Vidal-Beaudet, L.; Barthès, B.G. Quantification of soil organic carbon stock in urban soils using visible and near infrared reflectance spectroscopy (VNIRS) in situ or in laboratory conditions. *Sci. Total Environ.* **2019**, *686*, 764–773. [CrossRef] [PubMed]
13. Xie, S.; Li, Y.; Wang, X.; Liu, Z.; Ma, K.; Ding, L. Research on estimation models of the spectral characteristics of soil organic matter based on the soil particle size. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2021**, *260*, 119963. [CrossRef] [PubMed]
14. Yang, Y.; Shang, K.; Xu, Y. Analysis of Sensitive Spectral Characteristics of Farmland Soil Organic Matter Content Based on Ahsi/zy1-02d Data. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Brussels, Belgium, 11–16 July 2021. [CrossRef]
15. Luo, C.; Zhang, X.; Meng, X.; Zhu, H.; Ni, C.; Chen, M.; Liu, H. Regional mapping of soil organic matter content using multitemporal synthetic Landsat 8 images in Google Earth Engine. *Catena* **2022**, *209*, 105842. [CrossRef]
16. De Santana, F.B.; de Souza, A.M.; Poppi, R.J. Green methodology for soil organic matter analysis using a national near infrared spectral library in tandem with learning machine. *Sci. Total Environ.* **2019**, *658*, 895–900. [CrossRef]

17. Knox, N.M.; Grunwald, S.; McDowell, M.L.; Bruland, G.L.; Myers, D.B.; Harris, W.G. Modelling soil carbon fractions with visible near-infrared (VNIR) and mid-infrared (MIR) spectroscopy. *Geoderma* **2015**, *239–240*, 229–239. [CrossRef]
18. Liu, H.; Bao, Y.; Meng, X.; Cui, Y.; Zhang, A.; Liu, Y.; Wang, D. Inversion of soil organic matter based on GF-5 images under different noise reduction methods. *Trans. Chin. Soc. Agric. Eng.* **2020**, *36*, 90–98. [CrossRef]
19. Gu, X.; Wang, Y.; Sun, Q.; Yang, G.; Zhang, C. Hyperspectral inversion of soil organic matter content in cultivated land based on wavelet transform. *Comput. Electron. Agric.* **2019**, *167*, 105053. [CrossRef]
20. Xie, S.; Ding, F.; Chen, S.; Wang, X.; Li, Y.; Ma, K. Prediction of soil organic matter content based on characteristic band selection method. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2022**, *273*, 120949. [CrossRef]
21. Zhai, M. Inversion of organic matter content in wetland soil based on Landsat 8 remote sensing image. *J. Vis. Commun. Image Represent.* **2019**, *64*, 102645. [CrossRef]
22. Dou, X.; Wang, X.; Liu, H.; Zhang, X.; Cui, Y. Prediction of soil organic matter using multi-temporal satellite images in the Songnen Plain, China. *Geoderma* **2019**, *356*, 113896. [CrossRef]
23. Vaudour, E.; Gomez, C.; Lagacherie, P.; Loiseau, T.; Arrouays, D. Temporal mosaicking approaches of Sentinel-2 images for extending topsoil organic carbon content mapping in croplands. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *96*, 102277. [CrossRef]
24. Castaldi, F.; Palombo, A.; Santini, F.; Pascucci, S.; Pignatti, S.; Casa, R. Evaluation of the potential of the current and forthcoming multispectral and hyperspectral imagers to estimate soil texture and organic carbon. *Remote Sens. Environ.* **2016**, *179*, 54–65. [CrossRef]
25. Meng, X.; Bao, Y.; Liu, J.; Liu, H.; Kong, F. Regional soil organic carbon prediction model based on a discrete wavelet analysis of hyperspectral satellite data. *ITC J.* **2020**, *89*, 102111. [CrossRef]
26. Viscarra Rossel, R.A.; Cattle, S.R.; Ortega, A.; Fouad, Y. In situ measurements of soil colour, mineral composition and clay content by vis–NIR spectroscopy. *Geoderma* **2009**, *150*, 253–266. [CrossRef]
27. Dotto, A.C.; Dalmolin, R.S.D.; Caten, A.T.; Grunwald, S. A systematic study on the application of scatter-corrective and spectral-derivative preprocessing for multivariate prediction of soil organic carbon by Vis-NIR spectra. *Geoderma* **2018**, *314*, 262–274. [CrossRef]
28. Vašát, R.; Kodešová, R.; Klement, A.; Borůvka, L. Simple but efficient signal pre-processing in soil organic carbon spectroscopic estimation. *Geoderma* **2017**, *298*, 46–53. [CrossRef]
29. Zhang, Z.; Ding, J.; Zhu, C.; Wang, J. Combination of efficient signal pre-processing and optimal band combination algorithm to predict soil organic matter through visible and near-infrared spectra. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2020**, *240*, 118553. [CrossRef] [PubMed]
30. Swierenga, H.; Wülfert, F.; De Noord, O.E.; De Weijer, A.P.; Smilde, A.K.; Buydens, L.M.C. Development of robust calibration models in near infra-red spectrometric applications. *Anal. Chim. Acta* **2000**, *411*, 121–135. [CrossRef]
31. Galvão, R.K.H.; Araújo, M.C.U.; Silva, E.C.; José, G.E.; Soares, S.F.C.; Paiva, H.M. Short Report Cross-Validation for the Selection of Spectral Variables Using the Successive Projections Algorithm. *J. Braz. Chem. Soc.* **2007**, *18*, 1580–1584. [CrossRef]
32. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288. [CrossRef]
33. Wang, K.; Yang, S.; Guo, C.; Bian, X. Spectral Variable Selection Methods Based on LASSO Algorithm. *J. Instrum. Anal.* **2022**, *41*, 398–402. [CrossRef]
34. Li, D.; Chen, X.; Peng, Z.; Chen, S.; Chen, W.; Han, L.; Li, Y. Prediction of soil organic matter content in a litchi orchard of South China using spectral indices. *Soil Tillage Res.* **2012**, *123*, 78–86. [CrossRef]
35. Press, W.H.; Teukolsky, S.A. Savitzky-Golay Smoothing Filters. *Comput. Phys.* **1990**, *4*, 669. [CrossRef]
36. Tsai, F.; Philpot, W. Derivative Analysis of Hyperspectral Data. *Remote Sens. Environ.* **1998**, *66*, 41–51. [CrossRef]
37. Tsai, F.; Philpot, W.D. A derivative-aided hyperspectral image analysis system for land-cover classification. *IEEE Trans. Geosci. Remote Sens.* **2002**, *40*, 416–425. [CrossRef]
38. Minhoni, R.T.D.A.; Scudiero, E.; Zaccaria, D.; Saad, J.C.C. Multitemporal satellite imagery analysis for soil organic carbon assessment in an agricultural farm in southeastern Brazil. *Sci. Total Environ.* **2021**, *784*, 147216. [CrossRef]
39. Yu, L.; Hong, Y.; Gen, L.; Zhou, Y.; Zhu, Q.; Cao, J.; Nie, Y. Hyperspectral estimation of soil organic matter content based on partial least squares regression. *Trans. Chin. Soc. Agric. Eng.* **2015**, *31*, 103–109. [CrossRef]
40. Shen, L.; Gao, M.; Yan, J.; Li, Z.; Leng, P.; Yang, Q.; Duan, S. Hyperspectral Estimation of Soil Organic Matter Content using Different Spectral Preprocessing Techniques and PLSR Method. *Remote Sens.* **2020**, *12*, 1206. [CrossRef]
41. Demattê, J.A.M.; Ramirez-Lopez, L.; Marques, K.P.P.; Rodella, A.A. Chemometric soil analysis on the determination of specific bands for the detection of magnesium and potassium by spectroscopy. *Geoderma* **2017**, *288*, 8–22. [CrossRef]
42. Jakab, G.; Rieder, Á.; Vancsik, A.V.; Szalai, Z. Soil organic matter characterisation by photometric indices or photon correlation spectroscopy: Are they comparable? *Hung. Geogr. Bull.* **2018**, *67*, 109–120. [CrossRef]
43. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
44. Wang, X.; Zhang, F.; Kung, H.; Johnson, V.C. New methods for improving the remote sensing estimation of soil organic matter content (SOMC) in the Ebinur Lake Wetland National Nature Reserve (ELWNNR) in northwest China. *Remote Sens. Environ.* **2018**, *218*, 104–118. [CrossRef]
45. Yu, X.; Liu, Q.; Wang, Y.; Liu, X.; Liu, X. Evaluation of MLSR and PLSR for estimating soil element contents using visible/near-infrared spectroscopy in apple orchards on the Jiaodong peninsula. *Catena* **2016**, *137*, 340–349. [CrossRef]

46. Zhang, D.; Zhao, Y.; Qing, K.; Zhao, N.; Yang, Y. Influence of spectral transformation methods on nutrient content inversion accuracy by hyperspectral remote sensing in black soil. *Trans. Chin. Soc. Agric. Eng.* **2018**, *34*, 141–147. [CrossRef]
47. Dai, F.; Zhou, Q.; Lv, Z.; Wang, X.; Liu, G. Spatial prediction of soil organic matter content integrating artificial neural network and ordinary kriging in Tibetan Plateau. *Ecol. Indic.* **2014**, *45*, 184–194. [CrossRef]
48. Mehmood, T.; Liland, K.H.; Snipen, L.; Sæbø, S. A review of variable selection methods in Partial Least Squares Regression. *Chemom. Intell. Lab. Syst.* **2012**, *118*, 62–69. [CrossRef]
49. Vohland, M.; Ludwig, M.; Thiele-Bruhn, S.; Ludwig, B. Quantification of Soil Properties with Hyperspectral Data: Selecting Spectral Variables with Different Methods to Improve Accuracies and Analyze Prediction Mechanisms. *Remote Sens.* **2017**, *9*, 1103. [CrossRef]
50. Rossel, R.A.V.; Behrens, T. Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma* **2010**, *158*, 46–54. [CrossRef]
51. Galvao, L.S.I.S.; Vitorello, I. Variability of Laboratory Measured Soil Lines of Soils from Southeastern Brazil. *Remote Sens. Environ.* **1998**, *63*, 166–181. [CrossRef]