Modeling and Prediction of Soil Organic Matter Content Based on Visible-Near-Infrared Spectroscopy

Chunxu Li †, Jinghan Zhao †, Yaoxiang Li *, Yongbin Meng © and Zheyu Zhang

Abstract: In order to explore the ever-changing law of soil organic matter (SOM) content in the forest of the Greater Khingan Mountains, a prediction model of the SOM content with a high accuracy and stability has been developed based on visible near-infrared (VIS-NIR) technology and multiple regression analysis. A total of 105 soil samples were collected from Cuifeng forest farm in Jagdaqi City, Greater Khingan Mountains region, Heilongjiang Province, China. Five classical preprocessing algorithms, including Savitzky–Golay convolution smoothing (S-G smoothing), standard normal variate transformation (SNV), multiplicative scatter correction (MSC), first derivative, second derivative, and the combinations of the above five methods were applied to the raw spectra. Wavelengths were optimized with five methods of competitive adaptive reweighted sampling (CARS), successive projections algorithm (SPA), uninformative variable elimination (UVE), synergy interval partial least square (SiPLS), and their combinations, and PLS models were developed accordingly. The results showed that when S-G smoothing is combined with SNV or MSC, both preprocessing strategies can improve the performance of the model. The prediction accuracy of SiPLS-PLS model and SiPLS-UVE-PLS model for the SOM content is higher than for other models, with an R² of 0.9663 and 0.9221, RMSEC of 0.0645 and 0.0981, R² of 0.9408 and 0.9270, and RMSEV of 0.0615 and 0.0683, respectively. The pretreatment strategies and characteristic variable selection methods used in this study could significantly improve the model performance and predicting efficiency.

Keywords: VIS-NIRS; forest soil; SOM; UVE; SiPLS

1. Introduction

Forests and soil are important terrestrial ecosystems, they have the role of regulating climate, improving soil quality, protecting biodiversity, and mitigating geologic hazards [1–4]. Soil plays an important role in the material cycle, energy flow, and information transfer in forest ecosystems, providing the necessary nutrients required for plant growth [5–7]. Soil nutrients directly affect the distribution, growth, and yield of forest timber and have an important impact on the distribution pattern of plant species within forest communities, thus affecting both forest ecosystems and terrestrial ecosystems [8–11]. The content of soil organic matter (SOM) is an important indicator of soil nutrients, which plays an important role in improving soil quality, increasing soil productivity, and enhancing soil erosion resistance [12].

Visible-near-infrared spectroscopy (VIS-NIRS) is a green, eco-friendly, simple, and effective qualitative and quantitative spectroscopic analytical technology. Coupled with suitable chemo-metric methods, NIR has been successfully applied in many fields such as the petrochemical, agriculture, food, pharmaceutical, and traditional Chinese medicine industries [13]. VIS-NIRS is the absorption spectrum of frequency multiplication and frequency synthesis of different H-base groups [14,15]. The application of visible near infrared reflectance spectroscopy (VIS-NIRS) to characterize the physical and chemical...
information of soil samples has become a popular method to predict soil properties and components. As Vis-NIRS is sensitive to the organic components of soil and can be used for the simultaneous detection of multiple components without sample preparation, the use of Vis-NIRS to predict the SOM content is one of the common and promising techniques for soil quality assessment and carbon sinks. While traditional laboratory methods for SOM determination are time-consuming, labor-intensive, and limited, Vis-NIRS is simple, efficient, environmentally friendly, and non-destructive, with the rapid and accurate determination of organic matter content in a large number of samples at a regional scale.

Due to the disadvantages of VIS-NIRS, such as the serious spectral band overlap and week effective signal, researchers have conducted lots of research on VIS-NIRS predictive modeling by combining chemometric methods with multivariate analysis methods. The research shows that the sample set division method and spectral preprocessing method have a large impact on the modeling results. Yuan et al. [16] used sequential division, the Kennard–Stone method (KS), and sample set partitioning based on joint x–y distances (SPXY) to divide the sample set of corn samples, separately, and input the training and test sets into the support vector machine (SVM) model for comparison. They found the SPXY method had a higher prediction accuracy than other methods.

Systematic and random errors cannot be avoided due to the effects of inhomogeneity of particle size, variability of moisture content, soil density, and instrument noise [17]. Zhu et al. [18] built synergy interval partial least square (SiPLS), SVM, and random forest (RF) models for soil samples based on six spectral preprocessing methods, and the results showed that suitable preprocessing methods can improve model performance by eliminating noise and extracting effective information. In addition, wavelength variable extraction can also compress the input variables to a certain extent and improve the modeling efficiency. Guo et al. [19] applied four wavelength variable selection algorithms, namely synergy interval (SI), successive projections algorithm (SPA), genetic algorithm (GA), and competitive adaptive reweighted sampling (CARS) to reduce the dimensionality of the apple soluble solids spectra data, and the results showed that all methods can simplify the PLS model with correlation coefficients above 0.9, among which CARS had the best optimization effect.

In this study, a portable dispersive near infrared spectrometer (grating) is used for VIS-NIR spectrum acquisition. Our aims are modeling and predicting the SOM content based on VIS-NIRS and to examine the optimization algorithms in terms of spectral pretreatment and wavelength extraction. In this study, soil samples are collected at Cuifeng Forest Farm in northeast China and the SOM content is modeled and predicted with VIS-NIRS. Varied pretreatment algorithms are applied to the raw spectral of the soil and the better ones are selected for SOM content model development. The wavelength of the pretreated spectral is then optimized and the featured wavelength variables were extracted. A VIS-NIRS-based prediction model of the SOM content is then developed with the optimized wavelengths. Effective pretreatment methods and wavelength optimization algorithms significantly improve the model performance and aid in fast and accurate prediction of the SOM content.

2. Material and Methods
2.1. Soil Sample Collection

The sampling site is located at Cuifeng Forest Farm, Greater Khingan Mountains, Northeast China. The geographical coordinates are 123°45′45″~126°04′00″ E, 50°05′16″~51°12′40″ N. The geographical location is shown in Figure 1. It belongs to the low mountainous hilly area, with an elevation of about 430~520 m, gentle terrain, and a slope between 10°~15°. The study area has a cold-temperate continental monsoon climate with an annual average temperature of −1.2 °C, and uneven precipitation distribution with 70% of annual precipitation equal to 494.8 mm concentrated in June to August. The forest type is mainly natural secondary forest, and the dominant species
are Betula platyphylla Sukaczev, Larch, Mongolian oak, and mountain poplar. Its soil is dark brown loam forest soil with a thickness of about 20 cm.

Figure 1. Geographical location of Cuifeng Forest Farm.

In the study area, seven typical sample plots of 30 m × 30 m were selected from coniferous and broad-leaved mixed forests, each sample plot was divided into 15 6 × 10 m sub-plot. The soils were collected from each sub-plot by the five-point method [20], with a total of 105 soil samples collected. At each sub-plot, about 1 kg of soil from the depth of 10~20 cm in the soil profile was taken back to the laboratory for determination of the SOM content. Additionally, cutting rings with a volume of 100 cm³ were used to take soil samples at each sampling point in the sub-plot for determination of the physical properties of the soil, including bulk density, capillary porosity, and non-capillary porosity, etc. (Table 1). Statistical analysis was completed in IBM SPSS Statistics 26.0 (IBM, Armonk, NY, USA).

Table 1. Soil physical properties in the study area.

|                          | Absolute Water Content (%) | Bulk Density (g/cm³) | Saturated Water Capacity (%) | Capillary Water Capacity (%) | Non-Capillary Porosity (%) | Capillary Porosity (%) | Total Porosity (%) |
|--------------------------|----------------------------|----------------------|------------------------------|-----------------------------|----------------------------|----------------------|-------------------|
| Average                  | 31.61                      | 0.76                 | 80.82                        | 59.92                       | 15.58                      | 44.93                | 60.52             |
| SD                       | 8.00                       | 0.09                 | 10.47                        | 9.03                        | 3.70                       | 5.57                 | 4.88              |
| Max                      | 49.53                      | 1.03                 | 97.54                        | 75.95                       | 24.58                      | 54.08                | 68.54             |
| Min                      | 7.83                       | 0.60                 | 62.23                        | 43.25                       | 8.24                       | 31.92                | 49.10             |

Soil samples were air-dried in the laboratory to remove impurities, were ground and sieved (60 mesh), and the SOM content was determined by the wet oxidation method. The soil organic carbon (SOC) was completely oxidized to CO₂ using potassium dichromate (K₂Cr₂O₇) standard solution and concentrated sulfuric acid (chemically pure), and the
remaining K₂Cr₂O₇ was titrated with ferrous sulfate (FeSO₄) standard solution, the amount of FeSO₄ was recorded and the SOC content was calculated, and then the SOM content was obtained indirectly through conversion. The specific operation steps were guided by GB 7857-87 “Determination of organic matter in forest soils and calculation of carbon to nitrogen ratio”.

2.2. Visible-Near-Infrared Spectra Collection

The VIS-NIR spectra were collected using a LabSpec Pro FR/A114260 (Analytical Spectral Devices, Inc., Boulder, CO, USA). The wavelengths ranged from 350 to 2500 nm, with a minimum interval range of 2 nm, and a total of 2151 wavelength points were collected. The sampling time was 10 times/s. In order to eliminate the noise signal and improve the accuracy of spectral acquisition, the number of scans was increased to 30 times, and the average value of the repeated measurement spectra was taken as the original spectra of the soil samples, and the spectral signal-to-noise ratio was subsequently improved. The raw spectra of 105 soil samples were imported into the spectral data processing software ViewSpecPro to realize the conversion of diffuse reflectance into absorbance values according to Kubelka–Munk theory.

2.3. Modeling and Optimization

MATLAB R2016a (MathWorks, Natick, MA, USA) was used in this study for model development and the corresponding calculations.

The soil samples’ spectra were partitioned into calibration set and validation set based on joint x–y distances (SPXY). SPXY is derived from the classical sample classification algorithm Kennard–Stone method (KS), which first selects the two samples with the farthest Euclidean distances into the calibration set, then selects the samples with the largest and smallest distances into the calibration set by calculating the Euclidean distances from each remaining sample to each known sample in the calibration set, and so on, until the number of samples in the calibration set reaches the specified number [21]. SPXY is similar to the selection process of KS, with the difference being that SPXY integrates the spectral response matrix and concentration column vectors of the samples to achieve effective coverage of the multidimensional vector space, increase the variability and representativeness of the samples, and improve the stability of the established model [22]. The distance calculation equation of SPXY dividing the sample set is as follows:

\[
\begin{align*}
    d_x(p, q) & = \sqrt{\sum_{j=1}^{N} [x_p(j) - x_q(j)]^2} ; p, q \in [1,N] \\
    d_y(p, q) & = \sqrt{(y_p - y_q)^2} = |y_p - y_q| ; p, q \in [1,N] \\
    d_{xy}(p, q) & = \frac{d_x(p, q)}{\max_{p,q\in[1,N]} d_x(p, q)} + \frac{d_y(p, q)}{\max_{p,q\in[1,N]} d_y(p, q)} ; p, q \in [1,N]
\end{align*}
\]

where \(X_p\) and \(X_q\) denote two different samples and \(N\) is the number of wavelength points of the samples.

2.3.1. Spectral Preprocessing and Band Optimization

Pretreatment of the sample VIS-NIRS is aimed at eliminating systematic and random errors. Differential processing can reduce or eliminate some systematic errors, and methods such as smoothing can reduce random errors, so five classical preprocessing algorithms, namely, Savitzky–Golay convolution smoothing (S-G smoothing), standard normal variate transformation (SNV), multiplicative scatter correction (MSC), first derivative, and second derivative, were used. This experiment also explores the effect of combined strategy preprocessing to attenuate or even eliminate non-target information [23]. According to the combination of background correction and scattering correction [24], six combinations
of S-G smoothing-SNV, S-G smoothing-MSC, first derivative-SNV, first derivative-MSC, second derivative-SNV, and second derivative-MSC are used to further eliminate irrelevant information and noise. The effect of PLS modeling on the pretreatment spectra is evaluated in order to select the best pretreatment method.

VIS-NIRS are characterized by high dimensionality and redundancy, and have problems of high correlation, loss of rank, and covariance when applied to quantitative analysis, leading to poor accuracy and robustness of SOM content prediction models constructed based on the global spectra. In order to find the weak interference information that can characterize the samples and integrate multivariate information to eliminate the strong interference information, wavelength variable selection becomes a key issue in the applied analysis of SOM content of VIS-NIRS [25]. Wavelength variable selection can be divided into two types of point screening and band screening, and this experiment explored the effect of three wavelength point screening methods (CARS, SPA, and UVE) and one band screening method (SiPLS) on the improvement of model quality, and tried to couple different types of methods to screen the characteristic variables. In this experiment, the screening effects of six coupling strategies, UVE-CARS, UVE-SPA, CARS-SPA, SiPLS-UVE, SiPLS-CARS, and SiPLS-SPA, were investigated.

CARS based on the principle of “survival of the fittest” in evolutionary theory has the characteristics of fast computation and a high screening efficiency [26]. This experiment sets the number of Monte Carlo samples to 100 and selects the subset with the smallest RMSECV by 10-fold cross-validation to find the set of optimal solution variables [27].

The successive projections algorithm (SPA) is capable of extracting the few characteristic wavelengths in the full waveband that contain the least redundant information and the least covariance, thus eliminating a larger number of irrelevant variables in the original spectral matrix, and is widely used in the field of high-precision analysis of NIRS [28]. For the original eigenspectral matrix $X_{m \times n}$ (m is the number of samples and n is the number of wavelengths), $x_{m(0)}$ and $K$ are the initial iteration vector and the number of wavelengths to be extracted, respectively. SPA is a forward cyclic variable selection method that minimizes vector space covariance [29].

According to the relationship between the spectral matrix and the concentration matrix, the uninformed variable elimination method (UVE) was used to select characteristic wavelengths based on the regression coefficients of the PLS model [30]. Based on the PLS model, the randomly generated noise matrix with the same number of variables as the dependent variable matrix was introduced and combined with the original spectral matrix to form a new matrix. $C_i$ in the interval of $[1, n]$, which is less than $C_{max}$, is excluded and the remaining variables are extracted to form $X_{UVE}$. The specific formula is as follows:

$$C_i = \frac{\text{mean}(b_i)}{\text{std}(b_i)}, \quad (i = 1, 2, \cdots, 2n)$$  \hspace{1cm} (4)

SiPLS is an extension and improvement of the interval partial least square (iPLS) method. iPLS splits the full-band data into several equal-width intervals, builds a PLS model for each subinterval separately, and selects the optimal modeling interval using RMSECV as the model indicator [31,32]. However, iPLS is sensitive to the interval width and the selected subinterval may not be exactly the information interval. SiPLS overcomes the drawback of iPLS single interval modeling by calculating all possible PLS combination models for two or more subintervals in the same interval division based on the idea of permutation and combination, and the optimal joint subinterval is used for PLS regression.

Each of the wavelength selection methods have their limitations. In order to improve the reliability of the characteristic wavelengths, this experiment explored the band optimization effects of six combination strategies: UVE-CARS, UVE-SPA, UVE-SiPLS, CARS-SPA, SiPLS-CARS, and SiPLS-SPA [33,34].
2.3.2. Modeling Methods and Model Evaluation

Partial least square regression (PLSR) combines the advantages of principal component analysis (PCA), typical correlation analysis (CCA), and multiple linear regression (MLR) to achieve high-dimensional data structure simplification, regression modeling, and analysis of multiple correlations of independent variables [35]. The main idea of PLSR-based NIR prediction model for SOM is to reduce the dimensionality and reveal the main influencing factors of organic matter content variation from the spectral data with a much smaller number of samples than the dimensionality, so that the model is robust. A partial least square (PLS) model was established by using NIRS combined with the multivariate selection method to quantitatively analyze the SOM content in mixed coniferous low-quality forest in the Jagdaqi region. Coefficient of determination ($R^2$) and root mean square error (RMSE) were selected for the model evaluation. The coefficient of determination ($R^2$) is a concept in analysis of variance and regression analysis. It is a measure of the proportion of explained variance present in the data. The larger the $R^2$ is in the range of 0 to 1, the better the stability of the model, and the higher the fitting degree [36]. RMSE is used to detect the deviation between the estimated value and the true value. The smaller the RMSE, the better the prediction ability of the model. The computation equations for these criteria are as follows [37]:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST},$$

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

3. Results

3.1. Partition of Sample Sets

Based on SPXY, 105 soil samples were divided into the calibration set (70 soil samples) and validation set (35 soil samples), with a ratio of 2:1 (Table 2).

Table 2. Soil organic matter content in the study area.

| Plot No. | Number of Samples | Maximum (g/kg) | Minimum (g/kg) | Average Value (g/kg) | Standard Deviation (g/kg) |
|----------|-------------------|----------------|----------------|----------------------|--------------------------|
| 1        | 15                | 11.9713        | 8.1776         | 10.294               | 1.1384                   |
| 2        | 15                | 11.4979        | 6.8753         | 8.942               | 1.6408                   |
| 3        | 15                | 9.1034         | 4.7914         | 6.923               | 1.2043                   |
| 4        | 15                | 15.6771        | 5.9419         | 9.319               | 2.6095                   |
| 5        | 15                | 13.5342        | 7.4258         | 9.646               | 1.6054                   |
| 6        | 15                | 15.9396        | 6.7213         | 12.248              | 3.1646                   |
| 7        | 15                | 21.6961        | 7.3095         | 14.7311             | 4.3775                   |
| Total    | 105               | 21.6961        | 4.7914         | 10.3005             | 3.3790                   |

3.2. Spectral Preprocessing

Taking the original spectra as the control set, five pretreatment methods and six combination methods were investigated in this study in terms of the model performance (Figure 2). The results were compared with the original VIS-NIR spectral for model optimization and the best pretreatment methods were selected for VIS-NIR based SOM modeling and prediction. Figure 2A shows the original spectrum of the soil powder, and each color corresponds to a spectrum.
Comparing the S-G smoothing spectrum (Figure 2B) with the original spectrum shows that the S-G smoothing with a filter window width of 5 and a quadratic polynomial fit to the data points completely retains the effective information of the spectrum, and the shape of the spectrum tends to be smooth at this point. The effects of uneven solid particle size, surface scattering, and optical path changes on diffuse reflectance spectra were eliminated in the NIRS pretreated with SNV and MSC, and the spectral characteristic peaks were also highlighted. It can be seen from Figure 2C,D that the characteristic variables were mainly concentrated in the vicinity of 550–600 nm, 1000–1100 nm and 1400–1900 nm. The soil spectra after the first derivative and the second derivative processing are shown in Figure 2I–L, the combination of SNV and MSC correction of the first derivative and second derivative processing to extract the differences between adjacent wavelength variables amplified the noise contained in the original spectra, which cannot be effectively corrected for errors and has a lower signal-to-noise ratio [38].

The different combinations of preprocessing strategies follow the optimization window parameters used in the preprocessing of a single algorithm. S-G smoothing -SNV (Figure 2G) and S-G smoothing -MSC (Figure 2H) compared with SNV and MSC, the S-G smoothing followed by scattering correction results in significant differences in absorbance between different samples and stronger characteristics of multiplicity and ensemble. As shown in Figure 2I–L, the combination of SNV and MSC correction of the first derivative and second derivative processing to extract the differences between adjacent wavelength variables amplified the noise contained in the original spectra, which cannot be effectively corrected for errors and has a lower signal-to-noise ratio [38].

In the process of PLS modeling, the optimal number of latent variables (nLVs) needs to be determined by 10-fold cross-validation before further analysis of the coefficient of determination ($R^2$) and the root mean square error of prediction (RMSEP). The root mean square error of cross-validation (RMSECV) of the PLS model varies with the number of latent variables, and the selection of the optimal number of PLS latent variables follows the principle of minimum RMSECV. The number of principal components that minimizes or is almost invariant to the RMSECV is chosen under the premise that the number of components is as small as possible. The results of the 10-fold cross-validation of different pretreatment methods are shown in Figure 3. The RMSECV of the soil organic matter...
correction model established by original spectrum with five latent variables selected is the minimum. The optimal number of latent variables for the PLS correction model was 6, 4, 4, 1, and 1 after S-G smoothing, SNV, MSC, first derivative, and second derivative treatments, respectively. The optimal number of latent variables was 6, 6, 1, 1, and 2 after S-G-SNV, S-G-MSC, first derivative-SNV, first derivative-MSC, second derivative-SNV, and second derivative-MSC, respectively.

![Figure 3. The results of 10-fold cross-validation. (A) Original spectra; (B) S-G smoothing spectra; (C) SNV; (D) MSC; (E) 1st derivative; (F) 2nd derivative; (G) S-G-SNV; (H) S-G-MSC; (I) 1st derivative-SNV; (J) 1st derivative-MSC; (K) 2nd derivative-SNV; (L) 2nd derivative-MSC.](image)

The PLS modeling results of the SOM content under different pretreatment are shown in Table 3, with the $R^2 = 0.7781$ and RMSEP = 0.1655 of the original PLS model. The S-G-PLS calibration model has the best performance in a single preprocessing method, with $R^2 = 0.7968$ and RMSEP = 0.1584. Among the combined strategies, the S-G-SNV-PLS model and the S-G-MSC-PLS model outperformed the other models, with $R^2$ of 0.8082 and 0.8072, and RMSEP of 0.1539 and 0.1543, respectively.

| Sample Grouping        | Number of Samples | Minimum (g/kg) | Maximum (g/kg) | Average (g/kg) | Standard Deviation (g/kg) | Variance (g/kg) |
|------------------------|-------------------|----------------|----------------|----------------|---------------------------|-----------------|
| Calibration set        | 70                | 3.0344         | 21.6956        | 10.9455        | 3.6987                    | 1.3678          |
| Validation set         | 35                | 5.8685         | 20.8697        | 9.4811         | 3.5534                    | 1.2624          |

A comprehensive comparison of the PLS models with five single preprocessing and six different combinations of preprocessing shows that the optimal preprocessing methods are S-G-SNV and S-G-MSC. The PLS model with S-G smoothing outperforms other PLS models
with a single pre-treatment, indicating that S-G smoothing, either alone for NIRS or in combination with scattering correction, improves modeling to some extent compared to the original spectra. However, the overall noise reduction effect is weaker when S-G smoothing is performed alone. Therefore, it is necessary to preprocess the soil spectra using a suitable combination algorithm. After the NIRS data were preprocessed by a single algorithm, SNV had a stronger correction capability than MSC because MSC operates based on the spectral array of each sample group, while SNV operates on each spectrum. However, smoothing improves the model performance more than baseline correction and scattering correction. The PLS models of 1D-SNV, 1D-MSC, 2D-SNV, and 2D-MSC all performed lower than the original spectra and the single preprocessed spectra, indicating that preprocessing with two or more methods is not necessarily superior to a single preprocessing method. In this experiment, both the S-G-SNV-PLS correction model and the S-G-MSC-PLS correction model outperformed the other models, and the S-G-SNV and S-G-MSC treated soil full-spectrum data were used for variable screening.

3.3. Wavelength Selection

In this study, four single methods and six combined strategies were used to select wavelength variables for VIS-NIRS of SOM content, which are CARS, SPA, UVE, SiPLS, UVE-CARS, UVE-SPA, UVE-SiPLS, CARS-SPA, SiPLS-CARS, and SiPLS-SPA.

CARS screens the pretreated SOM content spectra of S-G-SNV and S-G-MSC, as shown in Figure 4. The S-G-SNV pretreatment spectra showed an overall decreasing trend of RMSECV although fluctuating when the sampling times was increased from 0 to 60, and its value reached the minimum at the 56th sampling. Then, 73 and 17 variables were selected separately for the calibration set and validation set, accounting for 3.39% and 0.79% of the original variables, respectively. The 52nd CARS sampling selected 63 and 14 variables, respectively, in the S-G-MSC calibration and validation set spectra, each representing 2.93% and 0.65% of the global variables. The S-G-SNV-CARS-PLS model worked better with the calibration model nLVs = 15, Rv2 = 0.9520, and RMSEC = 0.0770; the validation model nLVs = 14, Rv2 = 0.9120, and RMSEV = 0.0750.

Figure 4. Variable selection results of SOM content VIS-NIRS based on CARS. (a) Effect of CARS on S-G-SNV pretreatment spectra and (b) effect of CARS on S-G-MSC pretreatment spectra.

The accuracy and stability of the established PLS models vary greatly when SPA extracts different numbers of wavelength points [39]. S-G-SNV and S-G-MSC pretreatment both minimize the RMSE when 30 feature wavelengths were extracted, with RMSE of 0.2602 and 0.2329. The extraction of feature wavelengths is shown in Figure 5, where the red quadrilateral is used to mark the specific positions of the feature wavelengths.
The analysis shows that the feature wavelengths extracted by SPA overlap well in two different pretreatment backgrounds, especially those near 2200–2500 nm. For instance, 2194 nm are highly correlated with ArCH aromatic hydroxyl groups, and 2227, 2271, 2307, 2330, 2343, 2452, 2476 and 2481 nm are correlated with methyl groups (CH2), methylene group (CH1) and hypomethyl group (CH) [39]. The characteristic wavelengths which accounted for 1.39% of the global variables, were entered into the SOM content PLS model in order of importance. And the optimal number of latent variables corrected and validated by both the S-G-SNV-SPA-PLS model and the S-G-MSC-SPA-PLS model were consistent. The S-G-SNV-CARS-PLS model worked better with the calibration model \( R_c^2 = 0.7896, \) RMSEC = 0.1612; the validation model \( R_v^2 = 0.8588 \) and RMSEV = 0.0950.

The effect of UVE screening is shown in Figure 6, where the left side of the vertical dividing line shows the original wavelength variables, the right side of the red area shows the randomly introduced noise variables, and the upper and lower solid lines parallel to the horizontal axis are the threshold lines for assessing the stability. From the principle of UVE, it can be seen that the wavelength points between the threshold lines do not contribute to the modeling and are uninformative variables. The parts marked with green “*” beyond the threshold are the feature variables containing the modeling information. The number of variables remaining after the UVE screening of S-G-SNV and S-G-MSC preprocessed spectra are 52 and 53, each accounting for 2.42% and 2.46%, respectively, of the global variables. Under the condition that the number of latent variables was 14, the S-G-MSC-UVE-PLS model was slightly more effective than the PLS model constructed after S-G-SNV treatment, with \( R_c^2 = 0.8173, R_v^2 = 0.7580, \) RMSEC = 0.1502, and RMSEV = 0.1243.

This study explored the effect of the number of interval combinations on the performance of the SOM content PLS models. The effect of SiPLS on the band preference of the different pretreatment spectra when the number of interval combinations was 2, 3, and 4 is shown in Table 4 [40–42]. The results show that the optimal number of intervals was 3. The RMSECV of both types of SiPLS models tends to decrease when the number of selected intervals to be combined increased from 2 to 3. However, as the number of combinations increased to 4, more of the input variables led to noise being introduced into the model, which increased the error and reduced the prediction accuracy. As can be seen from Figure 7, the PLS factor was 6 for both the S-G-SNV and S-G-MSC pretreated spectra, and the selected intervals were the same, 458–565 nm, 1430–1537 nm, and 1645–1751 nm, at which the RMSE obtained the minimum values of 0.2131 and 0.2120, respectively, when the number of interval combinations was equal to 3. After S-G-SNV pretreatment, the \( R_c^2 \)
and $R_v^2$ of the SiPLS-PLS model were 0.9663 and 0.9408, and the RMSEC and RMSEV were 0.0645 and 0.0615, respectively. The characteristic spectra treated with S-G-MSC showed $R_c^2 = 0.9659$, $R_v^2 = 0.9442$, RMSEC = 0.0649, and RMSEV = 0.0597.

![Figure 6](image)

Figure 6. Band screening results of SOM content VIS-NIRS based on UVE. (a) Variable combination of S-G-SNV pretreated spectra screened by UVE and (b) variable combination of S-G-MSC pretreated spectra screened by UVE.

| Spectral Preprocessing   | Calibration Models | Validation Models |
|--------------------------|--------------------|-------------------|
|                          | nLVs | $R^2$ | RMSEP | nLVs | $R^2$ | RMSEP |
| Original Spectrum        | 5    | 0.7781 | 0.1655 | 2    | 0.4992 | 0.1789 |
| S-G Smoothing            | 6    | 0.7968 | 0.1584 | 2    | 0.4284 | 0.1911 |
| SNV                      | 4    | 0.7640 | 0.1707 | 1    | 0.4142 | 0.1935 |
| MSC                      | 4    | 0.7620 | 0.1715 | 1    | 0.4131 | 0.1936 |
| 1st Derivative           | 1    | 0.2541 | 0.3078 | 6    | 0.9699 | 0.0438 |
| 2nd Derivative           | 1    | 0.2332 | 0.3079 | 1    | 0.4781 | 0.1826 |
| S-G-SNV                  | 6    | 0.8082 | 0.1539 | 1    | 0.3603 | 0.2022 |
| S-G-MSC                  | 6    | 0.8072 | 0.1543 | 1    | 0.3595 | 0.2023 |
| 1st Derivative -SNV      | 1    | 0.3018 | 0.2937 | 1    | 0.4930 | 0.1800 |
| 1st Derivative -MSC      | 1    | 0.3018 | 0.2937 | 1    | 0.4930 | 0.1800 |
| 2nd Derivative -SNV      | 1    | 0.2580 | 0.3027 | 1    | 0.4603 | 0.1857 |
| 2nd Derivative -MSC      | 2    | 0.3227 | 0.2892 | 9    | 0.7670 | 0.1220 |

Different wavelength selection methods have their limitations. To further improve the reliability of the characteristic wavelength, on the one hand, multiple variable selection methods can be used to further refine the classical algorithm and make up for the defects of the classical algorithm. On the other hand, the noise and non-informative intervals can be eliminated by using the interval selection algorithm, and then other variable selection methods can be used to refine the selection. When the two variable selection methods are coupled, the optimization effect is not the same due to the different principles of wavelength selection between these two methods [33].

To compensate for the weakness of UVE in eliminating invalid variables, CARS was used to further refine the UVE, where 16 spectral feature variables of the S-G-SNV preprocessed correction set and 23 spectral feature variables of the S-G-MSC preprocessed correction set were selected by UVE-CARS, accounting for 30.77% and 43.40% of the UVE alone, respectively, which served to streamline the model through the input variables. The UVE-CARS-PLS models under the two preprocessing methods were established and compared with the UVE-PLS model, and the results are shown in Table 5. It was found that the combined screening did not affect the number of latent variables selected for
the calibration model, but the model accuracy and stability were improved due to the optimization of the input variables, with $R^2$ increasing to 0.8344 and 0.8735; $R^2_v$ increasing to 0.8503 and 0.8990; RMSEC reduced to 0.1430 and 0.1250; and RMSEV reduced to 0.0978 and 0.0803, respectively.

Figure 7. Interval combination results of SOM content VIS-NIRS based on SiPLS. (a) Synergy intervals of S-G-SNV pretreatment spectra and (b) synergy intervals of S-G-MSC pretreatment spectra.

Table 5. Band optimization results of two pretreated spectra by SiPLS.

| Spectral Preprocessing | Number of Combinations | Number of Selected Variables | Percentage of the Original Variables | Intervals | PLS Components | RMSE |
|------------------------|------------------------|------------------------------|--------------------------------------|-----------|----------------|------|
| S-G-SNV                | 2                      | 215                          | 10.00%                               | 2, 14     | 6              | 0.2249 |
|                        | 3                      | 323                          | 15.02%                               | 2, 11, 13 | 6              | 0.2131 |
|                        | 4                      | 430                          | 19.99%                               | 2, 9, 12, 13 | 6        | 0.2142 |
| S-G-MSC                | 2                      | 215                          | 10.00%                               | 2, 14     | 6              | 0.2224 |
|                        | 3                      | 323                          | 15.02%                               | 2, 11, 13 | 6              | 0.2120 |
|                        | 4                      | 430                          | 19.99%                               | 2, 9, 12, 13 | 6        | 0.2123 |

The number of variables selected for SiPLS-UVE increased compared to UVE. A total of 117 variables were selected for three subintervals in the S-G-SNV pretreatment context, while for S-G-MSC treated NIRs, the number of variables selected was 126. UVE coupled with SiPLS largely optimized the UVE-PLS model, and the optimal number of latent variables established by cross-testing was consistent with the results when SiPLS was performed independently, with both nLVs being 20. Modeled after S-G-SNV pretreatment, the SiPLS-UVE-PLS model worked slightly better than the S-G-MSC pretreatment, with the metrics shown in Table 5, where $R^2 = 0.9221$, $R^2_v = 0.9270$, RMSEC = 0.0981, and RMSEV = 0.0683.

The model complexity was significantly reduced by first collecting valid information intervals by SiPLS and then filtering out redundant variables by CARS. 30 and 12 feature variables each were involved in the construction of the SiPLS-CARS-PLS model after the combined S-G-SNV strategy treatment for its calibration and validation models, respectively accounting for 1.39% and 0.56% of the global variables. Similar results were obtained for the variable selection of the S-G-MSC pretreatment spectra, with 27 (1.26%) as well as 13 (0.60%) wavelength variables selected for each of the calibration and validation sets. The calibration model built on the basis of the 18 column vectors with high scores when
S-G-SNV pretreatment was combined with SiPLS-CARS band preference. On the one hand, the $R^2 = 0.9391$, RMSEC = 0.0867 of its calibration model. On the other hand, its validation model $R^2$ is 0.8401 and RMSEV is 0.1011.

SPA was used in combination with UVE, CARS and SiPLS to optimize the wavelength variables of the pretreatment spectrum. The three combinations all extracted 30 characteristic wavelengths, but the stability of the PLS model was significantly different. Cars-spa-pls model had the best performance among the three combined models. The latent variables of CARS-SPA-PLS model after S-G-SNV pretreatment were 15, $R^2 = 0.9520$, $R^2 = 0.8538$, RMSEC = 0.0770, RMSEV = 0.0967. After S-G -MSC pretreatment, the optimal number of latent variables of CARS-SPA-PLS model remained unchanged, $R^2$ and $R^2$ were 0.9524 and 0.8281, RMSEC and RMSEV were 0.0767 and 0.1048, respectively. The performance of the SiPLS-SPA-PLS model was between the UVE-SPA-PLS and CARS-SPA-PLS models. SiPLS-SPA selected more than 90% fewer variables than SiPLS, which greatly refined the results of the single SiPLS method to select feature variables. The UVE-SPA-PLS model had the worst selection effect of the wavelength variables, and compared with performing UVE alone, the percentage of the original variables decreased by 42.56% and 43.50%, respectively.

The results of the wavelength variable selection for a comprehensive comparison of the SOM content are shown in Table 6. The effects of SPA and UVE in the single wavelength variable selection method were relatively poor, and CARS and SiPLS had relatively good effects. The correction set and verification set of $R^2$ for the two preprocessing spectra of SIPLS were greater than 0.9 and the RMSE was less than 0.1. For CARS-PLS only under the preprocessing spectrum of SG-SNV, the $R^2$ of correction set and verification set were greater than 0.9, and the RMSE was less than 0.1. Among the six different combination strategies, only the $R^2$ of the correction set and verification set of SiPLS-UVE-PLS were greater than 0.9, and the RMSE of the SiPLS-UVE-PLS model pretreated by SG-SNV was less than 0.1. Through the comprehensive evaluation of the preprocessing method and wavelength selection method, the optimal PLS models were SiPLS-PLS and SiPLS-UVE-PLS after SG-SNV preprocessing.

### Table 6. PLS modeling results of different variable selection methods.

| Spectral Preprocessing | Feature Variable Selection Methods | Calibration Models | Validation Models |
|------------------------|-----------------------------------|--------------------|-------------------|
|                        |                                   | nLVs  | $R^2$ | RMSE | nLVs  | $R^2$ | RMSE |
| S-G-SNV                | -                                 | 6     | 0.8082 | 0.1539 | 1     | 0.3603 | 0.2022 |
|                        | CARS                              | 15    | 0.9520 | 0.0770 | 14    | 0.9120 | 0.0750 |
|                        | SPA                               | 22    | 0.7840 | 0.1633 | 22    | 0.8214 | 0.1068 |
|                        | UVE                               | 15    | 0.7964 | 0.1586 | 15    | 0.8092 | 0.1104 |
|                        | SiPLS                             | 20    | 0.9663 | 0.0645 | 20    | 0.9408 | 0.0615 |
|                        | UVE-CARS                          | 15    | 0.8344 | 0.1430 | 9     | 0.8503 | 0.0978 |
|                        | UVE-SPA                           | 7     | 0.7954 | 0.1590 | 7     | 0.6321 | 0.1533 |
|                        | SiPLS-UEVE                        | 20    | 0.9221 | 0.0981 | 20    | 0.9270 | 0.0683 |
|                        | CARS-SPA                          | 15    | 0.9520 | 0.0770 | 15    | 0.8538 | 0.0967 |
|                        | SiPLS-CARS                        | 18    | 0.9391 | 0.0867 | 12    | 0.8401 | 0.1011 |
|                        | SiPLS-SPA                         | 23    | 0.8084 | 0.1538 | 30    | 0.8267 | 0.1052 |
| S-G-MSC                | -                                 | 6     | 0.8072 | 0.1543 | 1     | 0.3595 | 0.2023 |
|                        | CARS                              | 15    | 0.9524 | 0.0767 | 7     | 0.7914 | 0.1155 |
|                        | SPA                               | 22    | 0.7896 | 0.1612 | 22    | 0.8588 | 0.0950 |
|                        | UVE                               | 14    | 0.8173 | 0.1502 | 14    | 0.7580 | 0.1243 |
|                        | SiPLS                             | 20    | 0.9659 | 0.0649 | 20    | 0.9442 | 0.0597 |
|                        | UVE-CARS                          | 14    | 0.8735 | 0.1250 | 16    | 0.8990 | 0.0803 |
|                        | UVE-SPA                           | 7     | 0.7879 | 0.1619 | 7     | 0.5838 | 0.1631 |
|                        | SiPLS-UEVE                        | 20    | 0.9081 | 0.1065 | 20    | 0.9117 | 0.0751 |
|                        | CARS-SPA                          | 15    | 0.9524 | 0.0767 | 15    | 0.8281 | 0.1048 |
|                        | SiPLS-CARS                        | 17    | 0.9250 | 0.0963 | 10    | 0.8509 | 0.0976 |
|                        | SiPLS-SPA                         | 23    | 0.8414 | 0.1400 | 30    | 0.8848 | 0.0858 |
Directly using SiPLS or SiPLS coupled with UVE, the feature variables preferred by both algorithms were input into the SOM content prediction model, and the validation set was also input to evaluate the model. The specific PLS modeling results are shown in Figure 8. The results visually reflect the correlation between the measured and predicted values of SOM content, and the trend line of its validation model overlaps with the target trend line nearly 1:1, indicating that the SiPLS-PLS and UVE-SiPLS-PLS models can achieve an effective prediction of SOM content within the sampling area. The fit between the trend line of its validation model and the target trend line is also high, indicating that the above two types of models can guarantee good prediction robustness under the premise of model accuracy.

Figure 8. Prediction model of the SOM content based on VIS-NIR. (a) SiPLS-PLS prediction model of SOM content and (b) UVE-SiPLS-PLS prediction model of the SOM content.

4. Conclusions and Discussion

PLS models were constructed based on NIRS combined with various chemometric algorithms to quantify the organic matter composition of low-quality mixed coniferous forest soils in the Jagdaq region. To ensure that the calibration set samples were fully representative and all samples were uniformly distributed within each set, SPXY was used to divide the data sets.

In this study, the effects of preprocessing by five single methods, as well as six combined methods were discussed. The results show that not all preprocessing methods are effective at eliminating noise and reducing errors for spectral data of complex sample systems. Compared with the original spectral modeling, there are three preprocessing strategies to improve the performance of the calibration model: S-G smoothing, S-G-SNV, and S-G-MSC. S-G-SNV and S-G-MSC with \( R^2 \) greater than 0.8 were selected as the optimal preprocessing.

The full spectrum of SOM content processed by S-G-SNV and S-G-MSC was used for wavelength variable screening. In this study, we selected 11 methods for feature variable selection (global PLS, CARS-PLS, SPA-PLS, UVE-PLS, SiPLS-PLS, UVE-CARS-PLS, UVE-SPA-PLS, UVE-SiPLS-PLS, CARS-SPA-PLS, SiPLS-CARS-PLS, and SiPLS-SPA-PLS) and the results were compared accordingly, the best ones were selected for model development.

When the four classical algorithms were executed individually, CARS was better at screening strong information variables and SiPLS was better at determining band combinations. Although SiPLS has the longest computation time and a weak ability to simplify the model inputs, the PLS regression analysis was significant. CARS selects the wavelength of
two different pretreated VIS-NIR spectra. The differences in its computed speed, lower feature variable screening, and difference in the correction model were small, but with a larger difference in $R^2_v$, explaining that CARS does not fully consider the synergy between adjacent variables under different backgrounds, which is consistent with the research results of Li Pao [43].

When the feature variable selection methods are used in combination, SiPLS-UVE is the best wavelength selection strategy. The ability of UVE-SPA to extract effective wavelengths was the worst and did not have the effect of improving model accuracy and precision. SiPLS-CARS and SiPLS-SPA further refined the variables ultimately involved in modeling. UVE-CARS and CARS-SPA simplified and optimized the full-spectrum model to some extent. This shows that when the feature variable selection methods were used in combination, some of the combinations could refine the variable screening model and improve the model run speed, and some of the combinations could improve the model quality and enhance the model prediction ability. The model simplification usually occurred along with the improvement of the model prediction ability. In addition, the selection of feature variables in different spectral preprocessing contexts could also have an impact on the constructed models. Therefore, which preprocessing method is more beneficial for modeling needs to be analyzed and discussed in conjunction with the wavelength variable selection process.

The results show that the optimal models are SiPLS-PLS and SiPLS-UVE-PLS, but not all feature variable selection methods can find strong information variables and eliminate weak information variables and irrelevant information variables. SiPLS-PLS and SiPLS-UVE-PLS compress global variables, reduce model calculation time, and enhance model prediction capabilities. The finally constructed model has a high prediction accuracy and good robustness, which can realize the effective prediction of SOM content in the sampling area. The two methods can provide methodological assistance for the rapid and accurate prediction of SOM content in the same forest type, and can provide the theoretical basis for the implementation of key forest management technologies, thus promoting the research progress of SOM dynamic monitoring of low-quality coniferous and broad-leaved mixed forest.

Author Contributions: Conceptualization, Y.L. and C.L.; methodology, C.L. and J.Z.; software, J.Z. and C.L.; validation, J.Z. and C.L.; formal analysis, J.Z.; investigation and sample collection, C.L., Y.M. and Z.Z.; writing—original draft preparation, C.L. and J.Z.; writing—review and editing, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Fundamental Research Funds for the Central Universities, grant number 2572018AB21, the Applied Technology Research and Development Plan of Heilongjiang Province (GA19C006) and the Key Research and Development Plan of Heilongjiang Province (GA21C030).

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

References
1. Wang, B.; Ren, X.; Hu, W. Assessment of Forest Ecosystem Services Value in China. Sci. Silvae Sin. 2011, 47, 145–153. [CrossRef]
2. Li, X.-B.; Zhao, Y.-L. Forest Transition, Agricultural Land Marginalization and Ecological Restoration. China Popul. Resour. Environ. 2011, 10, 91–95. [CrossRef]
3. Schwartzman, S.; Alencar, A.; Zarin, H.; Santos Souza, A.P. Social movements and large-scale tropical forest protection on the Amazon frontier: Conservation from chaos. J. Environ. Dev. 2010, 19, 274–299. [CrossRef]
4. Bai, Y.; Ouyang, Z.Y.; Zheng, H.; Xu, W.H.; Jiang, B.; Fang, Y. Evaluation of the forest ecosystem services in Haihe River Basin, China. Acta Ecol. Sin. 2011, 31, 2029–2039.
5. Han, J. Soil quality indicators and evaluation methods. Beijing Agric. 2016, 645, 162–163. [CrossRef]
6. Han, M.; Dong, X.; Guan, H.; Zhang, Q. Effects of Soil Properties on Ecological Functions in Different Succession Stages of Natural Larix gmelinii Forest in Daxing’an Mountains. J. Northeast. For. Univ. 2019, 47, 50–54. [CrossRef]
7. Wang, M.; Chen, H.; Zhang, W.; Wang, K. Soil nutrients and stoichiometric ratios as affected by land use and lithology at county scale in a karst area, southwest China. Sci. Total Environ. 2018, 619, 1299–1307. [CrossRef] [PubMed]
8. Qu, H.; Dong, X.; Tang, G.; Zhang, T.; Ma, X.; Guan, H. Effects of Replanting Alterations of Betula platyphylla Low-quality Forest on Soil Nutrients in Daxing’an Mountains. *J. Northeast. For. Univ.* 2017, 45, 75–80. [CrossRef]

9. Baldrian, P.; Kolarík, M.; Stursová, M.; Kopecký, J.; Valašková, V.; Větěvský, T.; Žifčáková, L.; Šnajdr, J.; Rídl, J.; Vlček, Č. Active and total microbial communities in forest soil are largely different and highly stratified during decomposition. *ISME J.* 2012, 6, 248–258. [CrossRef] [PubMed]

10. Danielson, R.; Visser, S. Effects of forest soil acidification on ectomycorrhizal and vesicular–arbuscular mycorrhizal development. *New Phytol.* 1989, 112, 41–47. [CrossRef]

11. Ji, H.; Dong, X. Comprehensive Evaluation of Soil Fertility after Transformation of the Low-Quality Forest in the Daxing’an Mountains. *Sci. Silvae Sin.* 2012, 48, 117–123. [CrossRef]

12. Kotroczó, Z.; Veres, Z.; Fekete, I.; Krakomperger, Z.; Tóth, J.A.; Lajthka, K.; Tóthmérész, B. Soil enzyme activity in response to long-term organic matter manipulation. *Soil Biol. Biochem.* 2014, 70, 237–243. [CrossRef]

13. Meng, Y.; Zhang, Y.; Li, C.; Zhao, J.; Wang, Z.; Wang, C.; Li, Y. Prediction of the Carbon Content of Six Tree Species from Visible-Near-Infrared Spectroscopy. *Forests* 2021, 12, 1233. [CrossRef]

14. Li, Y.; Via, B.K.; Cheng, Q.; Zhao, J.; Li, Y. New Pretreatment Methods for Visible–Near-Infrared Calibration Modeling of Air-Dry Density of Ulmus pumila Wood. *For. Prod. J.* 2019, 69, 188–194. [CrossRef]

15. Li, Y.; Via, B.K.; Young, T.; Li, Y. Visible-near infrared spectroscopy and chemometric methods for wood density prediction and origin/species identification. *Forests* 2019, 10, 1078. [CrossRef]

16. Yuan, Y.; Wang, W.; Chu, X.; Xi, M.J. Selection of Characteristic Wavelengths Using SPA and Qualitative Discrimination of Mildew Degree of Corn Kernels Based on SVM. *Spectr. Spectr. Anal.* 2016, 36, 226–230. [CrossRef]

17. Silalahi, D.D.; Midi, H.; Arasan, J.; Mustafa, M.S.; Caliman, J.-P. Robust generalized multiplicative scatter correction algorithm on pretreatment of near infrared spectral data. *Vib. Spectrosc.* 2018, 97, 55–65. [CrossRef]

18. Zhu, J.; Liu, Y.; Wu, C.; Jin, J.; Lv, H.; Yang, S. Study on near-infrared spectroscopy model of soil organic carbon after biochar addition and its application. *Acta Ecol. Sin.* 2020, 40, 7430–7440. [CrossRef]

19. Guo, Z.; Wang, M.; Ageyukum, A.A.; Wu, J.; Chen, Q.; Zuo, M.; El-Seedi, H.R.; Tao, F.; Shi, J.; Ouyang, Q. Quantitative detection of apple watercore and soluble solids content by near infrared transmittance spectroscopy. *J. Food Eng.* 2020, 279, 109955. [CrossRef]

20. Yang, H.; Jin, F.; Guan, T.; Xu, H.; Hu, X.; Xie, Q. Short-term Effect of Partial Substitution of Inorganic Fertilizer with Organic Fertilizer on Soil Fertility and Fungal Communities in Greenhouse. *Acta Agric. Borali-Occident.* 2021, 30, 422–430. [CrossRef]

21. Yang, Z.; Xiao, H.; Zhang, L.; Feng, D.; Zhang, F.; Jiang, M.; Sui, Q.; Jia, L. Fast determination of oxide content in cement raw meal using NIR spectroscopy with the SPXY algorithm. *Anal. Methods* 2019, 11, 3936–3942. [CrossRef]

22. Chen, Y.; Qi, T.; Huang, Y.; Wan, Y.; Zhao, R.; Yuan, L.; Zhang, C.; Fei, T. Optimization method of calibration dataset for VIS-NIR spectral inversion model of soil organic matter content. *Trans. Chin. Soc. Agric. Eng.* 2017, 33, 107–114. [CrossRef]

23. Divuš, P.; Bian, X.; Wang, Z.; Liu, W. Study on the Selection of Spectral Preprocessing Methods. *Spectrosc. Spectr. Anal.* 2019, 39, 2800–2806. [CrossRef]

24. Gerretzen, J.; Szymańska, E.; Jansen, J.J.; Bart, J.; van Manen, H.-J.; van den Heuvel, E.R.; Buydens, L.M. Simple and effective way for data preprocessing selection based on design of experiments. *Anal. Chem.* 2015, 87, 12096–12103. [CrossRef]

25. Mei, C.; Chen, Y.; Yin, L.; Jiang, H.; Chen, X.; Ding, Y.; Liu, G. Wavelength Selection by siPLS-LASSO for NIR Spectroscopy and Its Application. *Spectr. Spectr. Anal.* 2018, 38, 436–440. [CrossRef]

26. Gan, L.; Sun, T.; Liu, J.; Liu, M. Double Pulse LIBS Combined with Variable Screening to Detect Procymidone Content. *Spectrosc. Spectr. Anal.* 2019, 39, 584–588. [CrossRef]

27. Huo, Y.Q.; Zhang, C.; Li, Y.H.; Zhi, W.T.; Zhang, J. Nondestructive detection for kiwifruit based on the hyperspectral technology and machine learning. *J. Chin. Agric. Mech.* 2019, 40, 71–77. [CrossRef]

28. Zhang, Y.; Ren, D.; Han, Y.; Li, J. Air target reference spectrum selection based on characteristic wavelengths extracted by successive projections algorithm. *Infrared Laser Eng.* 2021, 50, 232–242. [CrossRef]

29. Huang, P.; Li, Y.; Yu, Q.; Wang, K.; Yin, H.; Hou, D.; Zhang, G. Classification of Organic Contaminants in Water Distribution Systems Developed by SPA and Multi-Classification SVM Using UV-Vis Spectroscopy. *Spectrosc. Spectr. Anal.* 2020, 40, 2267–2272. [CrossRef]

30. Sun, T.; Wu, Y.; Liu, X.; Mo, X.; Liu, M. Detection of Chromium Content in Soybean Oil by Laser Induced Breakdown Spectroscopy and UVE Method. *Spectrosc. Spectr. Anal.* 2016, 36, 3341–3345. [CrossRef]

31. Miao, X.; Miao, Y.; Gong, H.; Tao, S.; Chen, Y.; Chen, Z. Determination of Moisture Content in Rice by Near Infrared Spectroscopy with Different Partial Least Squares. *J. Anal. Sci.* 2019, 35, 639–643. [CrossRef]

32. Yang, H.; Zhu, M. Study of Rapid Detection of Soil Organic Matter Based on Characteristic Wavelength Selection of Visible-near Infrared Spectra. *Infrared Laser Eng.* 2015, 36, 42–48. [CrossRef]

33. Jia, M.; Li, W.; Wang, K.; Zhou, C.; Cheng, T.; Tian, Y.; Zhu, Y.; Cao, W.; Yao, X. A newly developed method to extract the optimal hyperspectral feature for monitoring leaf biomass in wheat. *Comput. Electron. Agric.* 2019, 165, 104942. [CrossRef]

34. Chen, Y.; Ma, H.; Zhang, Q.; Zhang, S.; Chen, M.; Wu, Y. Comparison of several variable selection methods for quantitative analysis and monitoring of the Yangxinshi tablet process using near-infrared spectroscopy. *Infrared Phys. Technol.* 2020, 105, 103188. [CrossRef]

35. Shen, L.; Gao, M.; Yan, J.; Yao, Y. Estimation model of soil organic matter based on SVR and PLSR. *China Agric. Inf.* 2019, 31, 58–71. [CrossRef]
36. Di Bucchianico, A. Coefficient of determination ($R^2$). In Encyclopedia of Statistics in Quality and Reliability; Champ, C.W., Shepherd, D.K., Eds.; John Wiley & Sons, Ltd.: Chichester, UK, 2008. [CrossRef]

37. Hong, Y.; Yu, L.; Zhu, Y.; Li, S.; Guo, L.; Liu, J.; Nie, Y.; Zhou, Y. Using Orthogonal Signal Correction Algorithm Removing the Effects of Soil Moisture on Hyperspectral Reflectance to Estimate Soil Organic Matter. *Sci. Agric. Sin.* 2017, 50, 3766–3777. [CrossRef]

38. Zhang, Y.; Zhou, M. Methods for Data Process of Near Infrared Spectroscopy Analysis. *Infrared Technol.* 2007, 29, 345–348. [CrossRef]

39. Zhang, H.; Luo, W.; Liu, X.; He, Y. Measurement of Soil Organic Matter with Near Infrared Spectroscopy Combined with Genetic Algorithm and Successive Projection Algorithm. *Spectrosc. Spectr. Anal.* 2017, 37, 584–587. [CrossRef]

40. Yang, Z.; Xiao, H.; Zhang, L.; Feng, D.; Zhang, F.; Jiang, M.; Sui, Q.; Jia, L. Fast determination of oxides content in cement raw meal using NIR spectroscopy combined with synergy interval partial least square and different preprocessing methods. *Measurement* 2020, 149, 106990. [CrossRef]

41. Chen, Y.; Di, Y.; Tang, X.; Cui, X.; Gao, X.; Cao, J.; Li, S. Combination Weight COD Concentration Prediction Model Based on BiPLS and SiPLS. *Spectrosc. Spectr. Anal.* 2019, 39, 2176–2181. [CrossRef]

42. Cheng, B.; Chen, D.; Wu, X. Near Infrared Spectral Wavelength Selection Based on Moving Window-Iterative Genetic Algorithm Method. *Chin. J. Anal. Chem.* 2006, 34, 123–130. [CrossRef]

43. Li, P.; Zhou, J.; Jiang, L.; Liu, X.; Du, G. A Variable Selection Approach of Near Infrared Spectra Based on Window Competitive Adaptive Reweighted Sampling Strategy. *Spectrosc. Spectr. Anal.* 2019, 39, 1428–1432. [CrossRef]