water

MDPI

Article

Prediction of Soil Water Content Based on Hyperspectral Reflectance Combined with Competitive Adaptive Reweighted Sampling and Random Frog Feature Extraction and the Back-Propagation Artificial Neural Network Method

Shaomin Chen [1,2], Fangchuan Lou [1,2], Yunfei Tuo [3], Shuai Tan [1,2,*], Kailun Peng [1,2], Shuai Zhang [1,2] and Quanjiu Wang [4]

[1] Faculty of Modern Agricultural Engineering, Kunming University of Science and Technology, Kunming 650500, China; chenshaomin1989@163.com (S.C.); fangchuanlou@163.com (F.L.)

[2] Yunnan Provincial Field Scientific Observation and Research Station on Water-Soil-Crop System in Seasonal Arid Region, Kunming University of Science and Technology, Kunming 650500, China

[3] Ecology and Environment Department, Southwest Forestry University, Kunming 650224, China

[4] State Key Laboratory of Eco-Hydraulics in Northwest Arid Region, Xi'an University of Technology, Xi'an 710048, China; wquanjiu@163.com

* Correspondence: tans90@163.com

Abstract: The soil water content (SWC) is a critical factor in agricultural production. To achieve real-time and nondestructive monitoring of the SWC, an experiment was conducted to measure the hyperspectral reflectance of soil samples with varying levels of water content. The soil samples were divided into two parts, SWC higher than field capacity (super-θ_f) and SWC lower than field capacity (sub-θ_f), and the outliers were detected by Monte Carlo cross-validation (MCCV). The raw spectra were processed using Savitzky–Golay (SG) smoothing and then the spectral feature variable of SWC was extracted by using a combination of competitive adaptive reweighted sampling (CARS) and random frog (Rfrog). Based on the extracted feature variables, an extreme learning machine (ELM), a back-propagation artificial neural network (BPANN), and a support vector machine (SVM) were used to establish the prediction model. The results showed that the accuracy of retrieving the SWC using the same model was poor, under two conditions, i.e., SWC above and below θ_f, mainly due to the influence of the lower accuracy of the super-θ_f part. The number of feature variables extracted by the sub-θ_f and super-θ_f datasets were 25 and 18, respectively, accounting for 1.85% and 1.33% of the raw spectra, and the variables were widely distributed in the NIR range. Among the models, the best results were achieved by the BPANN model for both the sub-θ_f and the super-θ_f datasets; the R^{2}p, RMSEp, and RRMSE of the sub-θ_f samples were 0.941, 1.570%, and 6.685%, respectively. The R^{2}p, RMSEp, and RRMSE of the super-θ_f samples were 0.764, 1.479%, and 4.205%, respectively. This study demonstrates that the CARS–Rfrog–BPANN method was reliable for the prediction of SWC.

Keywords: soil water content; hyperspectral reflectance; remote sensing retrieval; variable extraction; machine learning

check for updates

Citation: Chen, S.; Lou, F.; Tuo, Y.; Tan, S.; Peng, K.; Zhang, S.; Wang, Q. Prediction of Soil Water Content Based on Hyperspectral Reflectance Combined with Competitive Adaptive Reweighted Sampling and Random Frog Feature Extraction and the Back-Propagation Artificial Neural Network Method. *Water* **2023**, *15*, 2726. https://doi.org/10.3390/w15152726

Academic Editor: Songhao Shang

Received: 29 June 2023
Revised: 21 July 2023
Accepted: 26 July 2023
Published: 28 July 2023

1. Introduction

Water plays an important role as a transmitter in the SPAC (Soil–Plant–Atmosphere Continuum) system, creating a unified, dynamic, and interconnected system of mutual feedback between the soil, plants, and atmosphere. The soil water content (SWC) is a crucial parameter of soil physicochemical properties and is one of the necessary conditions for soil to nurture life. It is also one of the nonconstant parameters in agricultural, ecological, hydrological, and other research fields [1]. The SWC has been listed as an essential climate variable by the Global Climate Observing System [2]. In the agricultural industry, the SWC has always been a very important indicator, mainly playing a role in decision-making for irrigation management, efficient water use, and yield prediction [3]. Therefore, achieving rapid and accurate monitoring of the SWC status has always been a key concern for scholars, and the results of such research will play an important role in agricultural production.

Different methods have been developed to measure the SWC, including the oven-drying method, the resistance method, and the tensiometer method. Among the multitude of methodologies, the oven-drying method provides the most precise measurement for SWC. However, most of these methods have certain limitations. For instance, the oven-drying method involves destructive sampling, a tedious process, and poor real-time data [4]. The resistance soil moisture sensor is influenced by factors such as air gap, soil salinity, temperature, and bulk density, and even requires specific calibration [5]. Tensiometers have limitations such as lag and susceptibility to soil temperature and salinity, and they also require regular manual monitoring and maintenance [6]. Furthermore, these traditional methods are limited to point-scale measurements and do not provide spatially representative results, making it challenging to meet the requirements of real-time, large-scale, dynamic moisture estimation for precision agriculture. The active heated fiber optics (AHFO) method has demonstrated the potential to continuously determine the SWC at the field scale [7–9]. However, poor mobility, high cost, and professional post-maintenance have limited the widespread application of AHFO. Therefore, achieving large-scale SWC determination in real-time with accuracy and continuity remains a challenging task.

Spectrum technology has emerged as a rapidly developing analytical technique in recent years owing to its non-destructiveness, accuracy, and speed. Due to the inevitable defects of traditional SWC methods in monitoring on the spatial scale, spectrum technology has become a research direction for many scholars in SWC monitoring [10,11]. In the early stages of SWC spectral retrieval research, the majority of scholars focused on the diagnosis of soil moisture deficiency (much lower than field capacity (θ_f)). Bowers and Hanks [12] discovered that soil reflectance decreased as soil water increased in bare ground, and the spectral reflectance curve could be altered by the soil water [13,14]. As research progressed, the situation in which SWC was higher than the θ_f was studied. Neema et al. [15] pointed out that soil spectral reflectance decreased with increasing SWC when the SWC was below the θ_f and increased with increasing SWC when the SWC exceeded a certain threshold value. Liu et al. [16] demonstrated that the threshold is usually greater than the θ_f. Previous remote sensing retrieval studies tended to focus on the SWC below θ_f [17], and there were few studies reported on remote sensing retrieval SWC above θ_f. However, in the realm of agricultural production, farmers may face many situations that lead to a high SWC, such as heavy rainfall, over-irrigation, and poor drainage. This can negatively impact crop growth, resulting in a reduction in crop yield and even total crop failure [18]. Therefore, it is also of practical importance to diagnose an SWC above the θ_f. However, as the reflection spectrum of soil is a process that reduces initially and subsequently increases with an increase in SWC, using the same model to invert the SWC under the two conditions of water content above and below θ_f may lead to poor accuracy.

Hyperspectral data contain thousands of bands, many of which are mixed with noise and interfering variables. Data preprocessing and feature extraction algorithms can reduce noise, remove interfering variables, and improve model prediction [19]. However, when only one method is used to extract feature variables, the stability might be poor, and too many variables may be located, which would make the prediction model too complex [20]. To address the deficiency with feature band extraction methods, different variable extraction methods were used, for example, uninformative variable elimination plus the successive projections algorithm (UVE–SPA). The UVE–SPA method can cause the correlation between feature variables and targets to be more significant, while also reducing the number of variables [21]. Xu et al. [22] employed competitive adaptive reweighted sampling plus the successive projections algorithm (CARS–SPA) method to extract variables, which simplified the modeling process and improved the prediction accuracy of potato dry matter. Different coupled feature extraction methods have been studied in some fields, but it remains to be investigated whether the method can effectively extract SWC-sensitive bands (the SWC of samples included both lower and higher θ_f) and whether the dimensionality of the hyperspectral data can be sufficiently reduced to simplify model building.

Bowers and Hanks [12] reported absorption bands for soil water at 1400, 1900, and 2200 nm of indoor soil spectral reflectance, and the SWC could be predicted from the feature band 1900 nm. However, 1400 and 1900 nm are in the water–air absorption band, which is difficult to apply outdoors. Sun et al. [23] analyzed the absorption spectrum of black soil in northeast China and observed a strong correlation between the soil absorption spectrum and the SWC. The maximum absorbance peak point was found at 1946 nm, and the prediction dataset R^2 of the one-dimensional linear regression model of SWC was greater than 0.95. However, the soil composition is complex and variable, and a simple linear model may not retrieve the SWC accurately. Relevant findings have revealed that the relationship between soil spectral reflectance and SWC in a large range was usually nonlinear [13,16,24–26].

Machine learning has been widely applied in various fields in recent decades because of its ability to learn and approximate complex nonlinear mappings. In particular, quantitative remote sensing in agriculture has become an active research area for machine learning applications. The establishment of spectral monitoring of SWC based on the machine learning method is also an important research field. Research on estimating the SWC in saline soils also indicates that machine learning methods have more advantages, for example, the support vector machine model had better overall fitting ability compared to the multiple linear regression and partial least squares regression models [27]. The study of the spectral estimation of SWC in different soils (sandy and loamy) demonstrated that the nonlinear method (back propagation artificial neural network, BPANN) can predict well in single-soil and mixed-soil samples with $R^2 > 0.8$ [28]. Previous studies have demonstrated that machine learning methods are capable of effectively handling the nonlinearity of soil reflectance and SWC. However, further research is required to determine which of the commonly used machine learning models is best suited for inverting the SWC.

Based on this, the aims of this study were to (1) divide the sample into two parts with θ_f as the threshold to establish models, (2) extract the SWC-sensitive bands using a combination of the competitive adaptive reweighted sampling (CARS) and random frog (Rfrog) algorithms and evaluate the effectiveness of this integrated approach for identifying SWC-sensitive bands, and (3) establish and compare the performance of machine learning methods (extreme learning machine, back-propagation artificial neural network, and support vector machine) to select the optimal model for SWC prediction.

2. Materials and Methods

2.1. Preparation of Soil Samples

In this study, soil samples with a certain range of water content were obtained in the laboratory. Red soil was used for the soil sample preparation (porosity, 61.65%; bulk density, 1.01 g/cm^3; clay, 20.03%; silt, 62.32%; sand: 17.65%), and the collected soil raw materials were air-dried, finely ground, cleared of impurities, and made into test soil by passing through a 2 mm size sieve to reduce the effect of soil particle diameter on the spectral determination. The prepared soil was packed into a disc with a 16 cm inner diameter and 1.7 cm height, with several small holes at the bottom. After this, the soil sample surface was leveled and then placed into a tray with a water depth of approximately 1 cm to be saturated. The disc was removed and placed on air-dried soil lined with filter paper to allow the water to drain out naturally. Soil samples with various water contents were obtained by controlling the duration of the water removal. This process of soil sample preparation can avoid the uneven surfaces in soil samples caused by adding water from above.

2.2. Remote Sensing Data

The hyperspectral reflectance of the soil samples was determined using an SR-2500 portable geophysical spectrometer (Spectral Evolution, Inc., 1 Canal St., Unit B-1, Lawrence, MA 01840 USA). The wavelength range of the instrument was 350–2500 nm, with a total of 2151 channels. The portable spectroradiometer was equipped with optical fiber with a length of 1.5 m and an 8° field of view (FOV). The sampling intervals were 1.5 nm @ 350–1000 nm and 6 nm @ 1000–2500 nm, and the instrument automatically interpolated the measurement results into 1 nm intervals. To obtain steady spectral data, we chose a clear and cloudless day between 10:00 and 14:00 local time when the solar altitude angle and light intensity were optimal. During sampling, the optical fiber was placed 15 cm above the soil sample in a vertically downward position to ensure that the FOV coverage did not exceed the disc range. The hyperspectral data were collected 10 times for each soil sample, and the average value was utilized as the hyperspectral reflectance to reduce random errors. The instrument was calibrated using a standard whiteboard before measurement, and the calibration process was repeated every 10 min.

2.3. Soil Water Content Determination

After collecting the hyperspectral data, the SWCs of the samples were determined by the drying method (Table 1). The Wilcox method was used to measure the field water capacity of the experiment soil, which was 31.63% (mass water content).

Table 1. The descriptive statistics of the soil water content of samples.

Sample Size	Max (%)	Min (%)	Mean (%)	CV (%)
139	47.88	13.48	28.52	26.97

Note: CV means the coefficient of variance of the dataset. The same applies subsequently.

2.4. Spectral Preprocessing

In this study, the soil spectral reflectance was analyzed only in the 350–1349 nm and 1451–1800 nm bands (a total of 1350 bands) due to the presence of a strong absorption band of water near 1400 nm and the presence of large signal noise for reflectance greater than 1800 nm. During the acquisition process, the sample spectra were frequently disturbed by stray light, baseline drift, and other factors, which had an impact on the final analysis results. Therefore, it was necessary to preprocess the raw spectra. Savitzky–Golay (SG, window width, 3; polynomial, 1) smoothing was utilized to preprocess the raw spectral data.

2.5. Elimination of the Outliers and Sample Data Division

The collection, processing, and analysis of soil samples might introduce a degree of error, particularly human measurement error, which could affect subsequent data analysis and modeling. Samples with errors are called outliers, and it is often necessary to re-measure or eliminate them to minimize their impact on the subsequent processing results. To address the issue of outliers in the samples, this study employed Monte Carlo cross-validation (MCCV) [29] to identify them. The MCCV could efficiently detect outliers in the spectral array by analyzing the sensitivity of the prediction error to anomalous samples.

In the present study, all data were processed centrally. A total of 1000 PLSR models with SWC as the dependent variable and raw spectrum as the independent variable were established using MCCV, with a ratio of randomly selected samples of 0.7. The prediction error of each sample in the model was calculated, and the mean (MEAN) and standard deviation (STD) of the prediction errors for each sample were determined. A scatter plot illustrating the MEAN–STD of the sample set was created. Finally, 2.5 times the average value of either the MEANs or the STDs was taken as the threshold. The complete flow of the MCCV is shown in Figure 1.

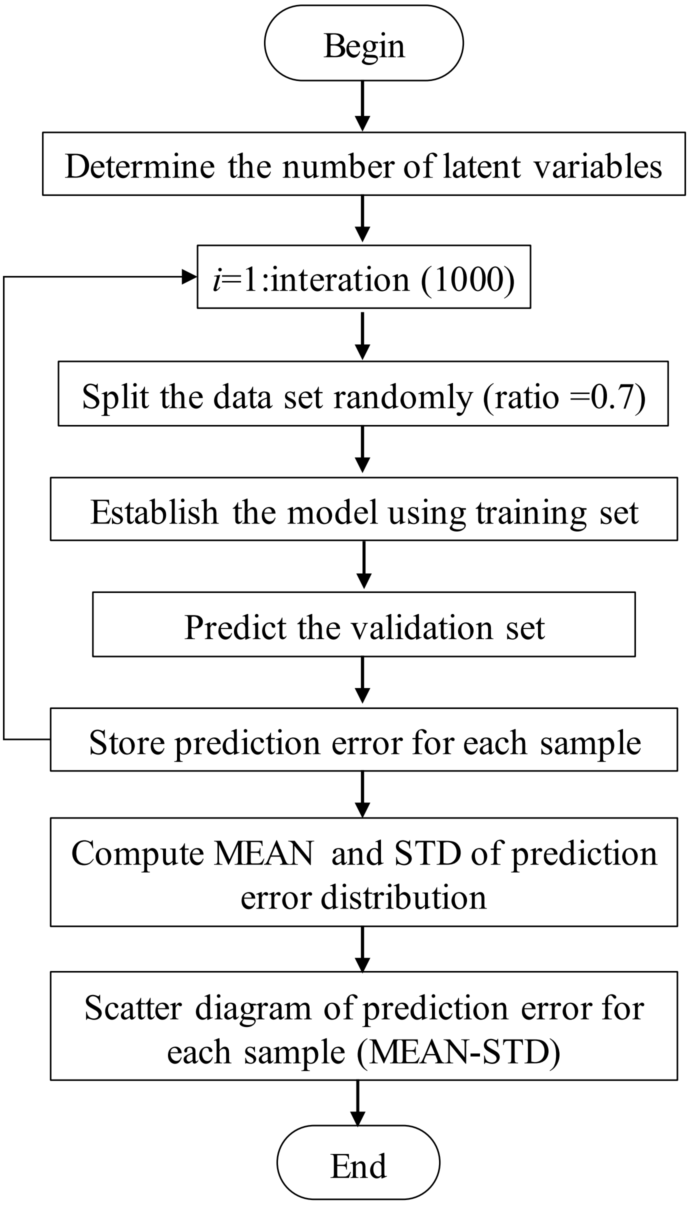
Begin
Determine the number of latent variables
i=1:interation (1000)
Split the data set randomly (ratio =0.7)
Establish the model using training set
Predict the validation set
Store prediction error for each sample
Compute MEAN and STD of prediction error distribution
Scatter diagram of prediction error for each sample (MEAN-STD)
End

Figure 1. The flow chart of MCCV.

During model construction, Sample Set Partitioning based on joint X–Y distance (SPXY) was implemented to divide the samples into representative calibration and prediction datasets with a ratio of 2:1. The SPXY algorithm, originally developed by Galvao [30], involved the calculation of the distance between each sample using spectral and target values as characteristic parameters to ensure difference and representativeness between the calibration and prediction datasets. This method effectively covered the multidimensional vector space and improved the model's prediction accuracy.

2.6. Feature Variable Extraction

The hyperspectral data contain a large amount of redundant data and irrelevant information in addition to information about the SWC, possibly leading to model complexity. Selecting important bands for the modeling not only reduces the complexity of the model but also results in better performance and higher accuracy.

In this study, competitive adaptive reweighted sampling (CARS) was chosen for spectral feature extraction [31]. The CARS algorithm mimicked the "survival of the fittest" principle of Darwinian evolutionary theory in selecting variables by treating wavelength variables as individual entities. During the selection process, bands with a strong adaptive capacity were retained, while those with a weak adaptive capacity were eliminated. As the CARS algorithm uses Monte Carlo sampling to randomly select modeling samples, the variable regression coefficients would change due to the random sample selection, and the absolute value of the regression coefficients cannot entirely indicate the significance of the variables, which affected the accuracy of the model.

To mitigate the influence of the randomness of the CARS algorithm, the random frog (Rfrog) algorithm was adopted to conduct a secondary data filtration after feature extraction by CARS, further simplifying the model while ensuring its accuracy. Rfrog is a feature selection algorithm proposed by Li [32], which operates iteratively. The variable selection process was executed using the reversible jump Markov chain Monte Carlo (RJMCMC) framework. A sufficient number ($\geq$10,000) of partial least squares regression (PLSR) models were built to calculate the selection probability of each band, and the probability of each band being selected was calculated in each iteration. The more information a band contains, the greater its selection probability. After completing the iterations, bands were ranked by their probability of being selected, and variables with a high probability of being selected were preferred as feature variables.

2.7. Modeling Method

In this study, based on the nonlinear characteristics between the soil spectral reflectance and the SWC [13,16], three nonlinear models, extreme learning machine (ELM), back-propagation artificial neural network (BPANN), and support vector machine (SVM), were selected for modeling.

The ELM is a single-hidden-layer feedforward neural network (SLFN) learning algorithm developed by Huang [33]. In contrast to conventional gradient-based feedforward neural network learning algorithms, the ELM randomly assigns weights and biases to the input layer. This algorithm's execution process may not require artificial parameter adjustment, avoiding repetitive iterations in the traditional training algorithm. As a result, the model trains extremely fast and achieves high generalization performance. In this work, the activation function of the hidden-layer neurons was set to "sigmoid" by default, and the number of hidden layers was initially set at 3, gradually increasing to 100 in steps of 1. Each model structure was operated multiple times to determine the optimal number of hidden-layer nodes based on the best results trained.

The BPANN is a widely used machine learning algorithm based on the gradient descent method, which uses gradient search techniques to reduce the mean squared error between the actual output value and the desired output value of the network. It consists of an input layer, a hidden layer, and an output layer, each containing several nodes. The weights of each node are calculated through self-learning to derive the training results. These results are analyzed for errors with the expected outcomes, and if the training results do not meet expectations, the weights are modified to reduce the errors. Continuous iteration helps to achieve consistency with the expected results and to minimize errors. The training function of the BPANN model was "newff"; the maximum iteration number was 10,000; the minimum error of the training target was 0.000001; the learning rate was 0.01; and the number of hidden-layer nodes was determined using the same method as for ELM.

The SVM is a learning system that uses linear function hypotheses in high-dimensional feature spaces [34]. Based on minimal structural risk, this method can better address practical problems such as the curse of dimensionality and overfitting. The proposed model effectively handles small samples, nonlinearity, high dimensions, and local minima and has good generalization ability. To better address the nonlinear characteristics of the data, the radial basis function (RBF) was used as SVM's kernel function in this study. There were two important parameters that needed to be adjusted in the model, i.e., the penalty factor (c) and the kernel function parameter (g). If either c or g is too large, the model prediction tends to be overfitted. By contrast, if either is too small, the model prediction tends to be underfitted. Either extreme situation could result in poor generalization ability.

A 5-fold cross-validation combined with the grid search method was used to find the optimal penalty factor c and kernel function parameter g within the range [2^{-10}, 2^{10}], with the step size $2^{0.5}$ to determine the final model.

2.8. Model Evaluation Metrics and Software

The coefficient of determination (R^2), root mean square error (RMSE), and relative root mean square error (RRMSE) were selected as the evaluation metrics. Generally, RRMSE > 10% represents that the model accuracy is excellent; furthermore, 10% < RRMSE < 30% represents that the model accuracy is good, and RRMSE > 30% represents that the model accuracy is poor.

The Unscrambler X 10.4 software was used for spectral preprocessing (SG smoothing). MATLAB 2020a was adopted for feature extraction and model building. Excel 2021 was employed for data analysis and scientific drawing.

3. Results

3.1. Spectral Preprocessing and Data Analysis

A prediction model was established by using partial least squares regression (PLSR) based on the full, pretreated spectrum (Figure 2). The scatter plot of parts with higher water content is more discrete than others, especially in the prediction dataset. Previous studies have shown that the soil spectral reflectance decreases as the SWC increases. However, when the SWC exceeds the θ_f of the soil, the soil spectral reflectance increases as the SWC increases.

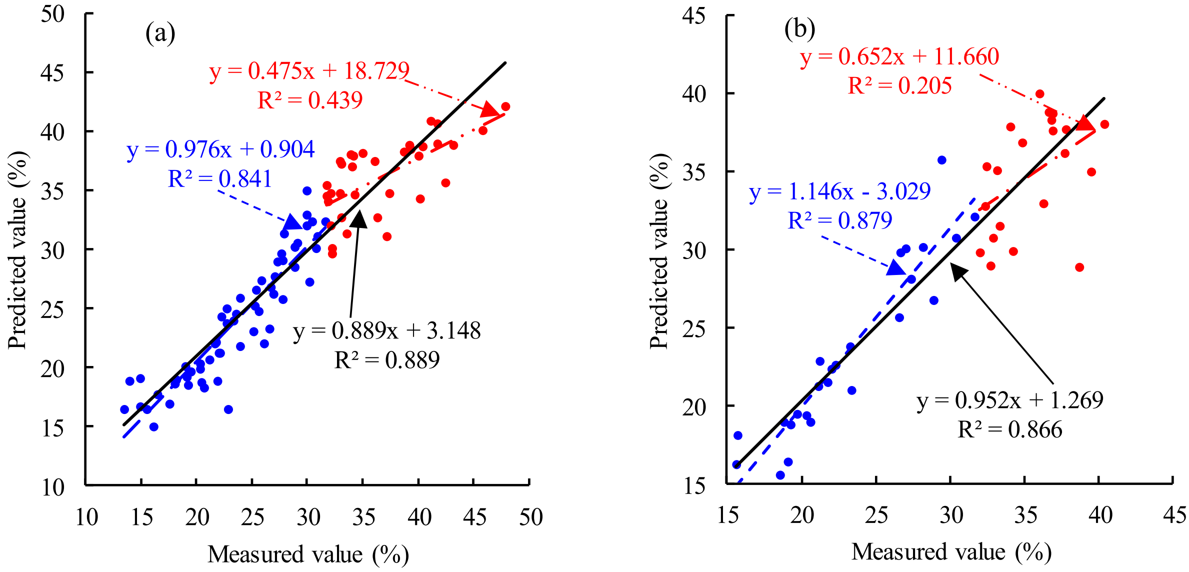

Figure 2. PLSR model based on the full, pretreated spectrum. (**a**) Calibration dataset of PLSR and (**b**) prediction dataset of PLSR. Blue scatter points represent sub-θ_f samples, while red scatter points represent super-θ_f samples; the black fitting line represents the total samples.

Considering this characteristic of soil reflectance, the results of the PLSR were divided into two parts: SWC higher than field capacity (super-θ_f) (red scatter points in Figure 2) and SWC lower than field capacity (sub-θ_f) (blue scatter points in Figure 2). Compared with the sub-θ_f samples, the accuracy of the model was lower in the super-θ_f samples with R^2 = 0.439 for calibration and R^2 = 0.205 for prediction. This indicates that combining the super-θ_f and sub-θ_f samples may be unreasonable to establish a model for the SWC prediction. To obtain a higher-accuracy model, the subsequent related work divided the soil samples into two parts: sub-θ_f contained 84 samples, and super-θ_f contained 55 samples. The two datasets were modeled independently.

3.2. Identification of the Outliers and Sample Division

The MCCV method was used to identify outliers; data centering was processed first, and the best latent variables were determined by PLSR. A set of predicted errors for each sample were obtained by using the MCCV method, and the MEAN and STD of the predicted errors were determined. A scatter plot of MEAN–STD was drawn (Figure 3), with 2.5 times the average value of MEAN and STD as the threshold (represented by the dotted line in Figure 3). The samples outside the dotted line were identified as outliers.

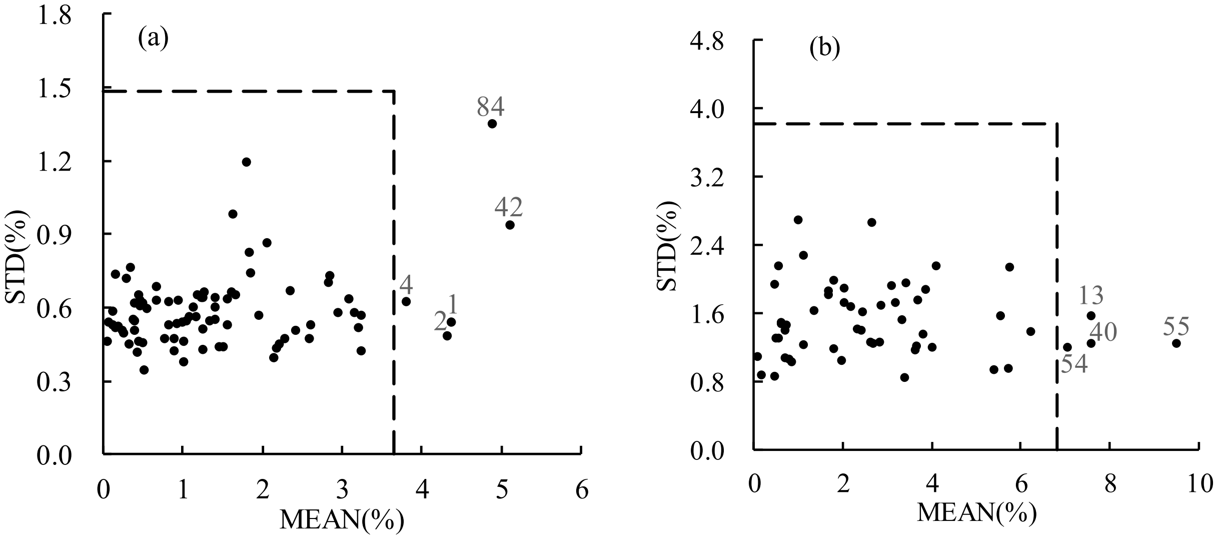

Figure 3. Results of outlier identification using MCCV. (**a**) Sub-θ_f samples, (**b**) super-θ_f samples.

As shown in Figure 3a, the sub-θ_f samples numbered 1, 2, 4, 42, and 84 were identified as outliers that should be eliminated, leaving 79 samples. As shown in Figure 3b, the super-θ_f samples numbered 13, 40, 54, and 55 were identified as outliers that should be eliminated, leaving 51 samples. The results of the descriptive statistics of the samples are displayed in Table 2. Based on the SPXY method, the calibration dataset and prediction dataset of the super-θ_f and sub-θ_f samples are shown in Table 3.

Table 2. Descriptive statistics of the samples divided by θ_f as the threshold.

Samples	Sample Size	Max (%)	Min (%)	Mean (%)	CV (%)
super-θ_f	51	43.13	31.72	35.87	9.29
sub-θ_f	79	31.59	14.90	23.71	18.91

Table 3. Descriptive statistics of the calibration dataset and prediction dataset.

Samples	Dataset	Sample Size	Max (%)	Min (%)	Mean (%)	CV (%)
super-θ_f	calibration dataset	34	43.13	31.72	36.25	10.04
	prediction dataset	17	40.36	31.79	35.11	7.24
sub-θ_f	calibration dataset	53	31.59	14.90	23.87	19.54
	prediction dataset	26	30.85	15.65	23.40	17.78

3.3. Feature Variable Extraction

For the feature variables extracted by the CARS method, the number of Monte Carlo sampling runs was set as 50 in the fivefold cross-validation, and all the data were processed centrally. As the number of sampling times increased, the number of extracted feature bands gradually decreased, eventually tending to zero (Figure 4a). During this process, the root mean square error of cross-validation (RMSECV) of the fivefold interaction validation PLSR model showed a trend of first decreasing and then increasing (Figures 4b and 5b).

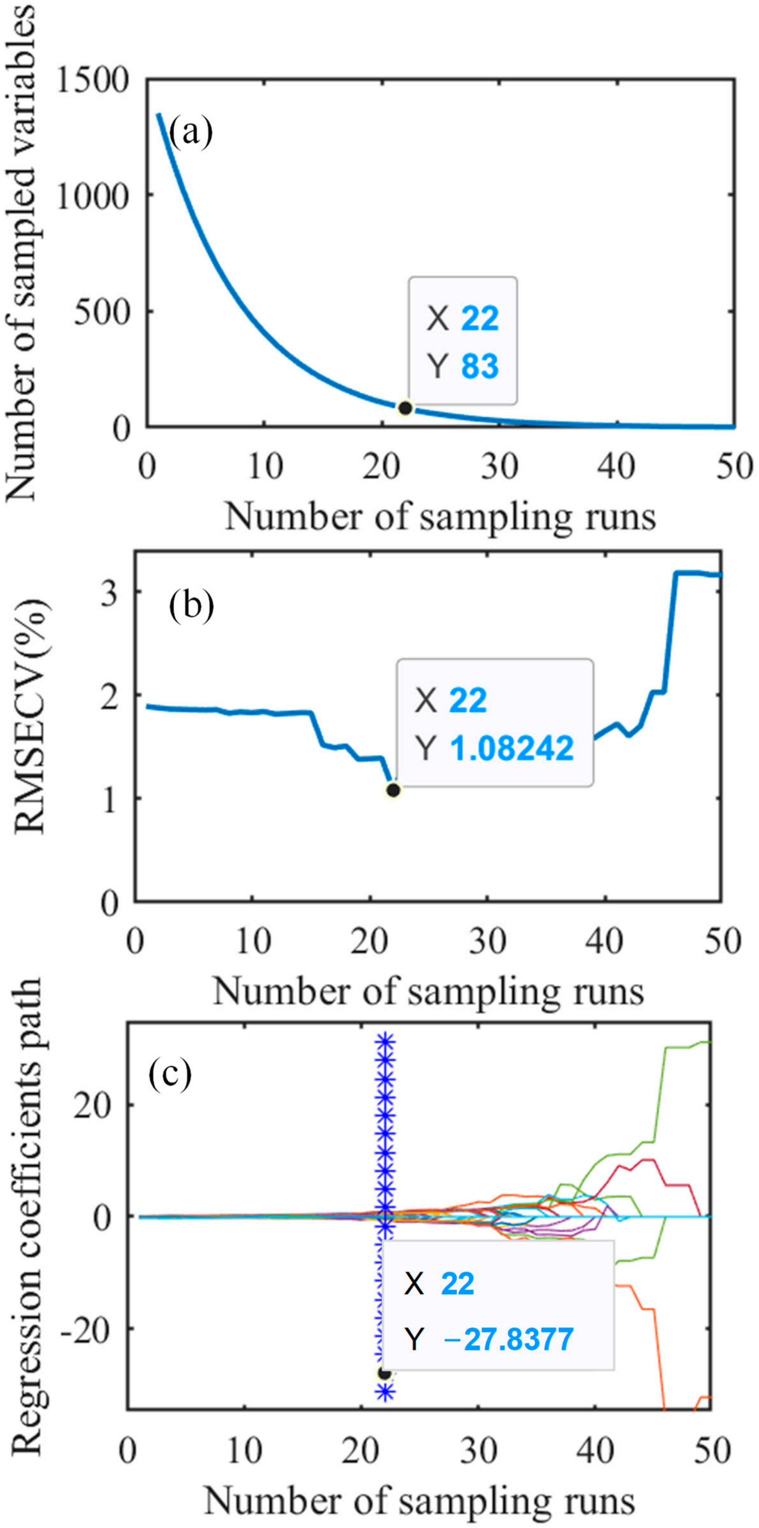

Figure 4. Results of CARS feature extraction for sub-θ_f samples. (**a**) Variable change trend, (**b**) Cross validation, (**c**) Trend of regression coefficient. The lines in different color showed the change trend of the regression coefficient of each wavelength with number of sampling runs.

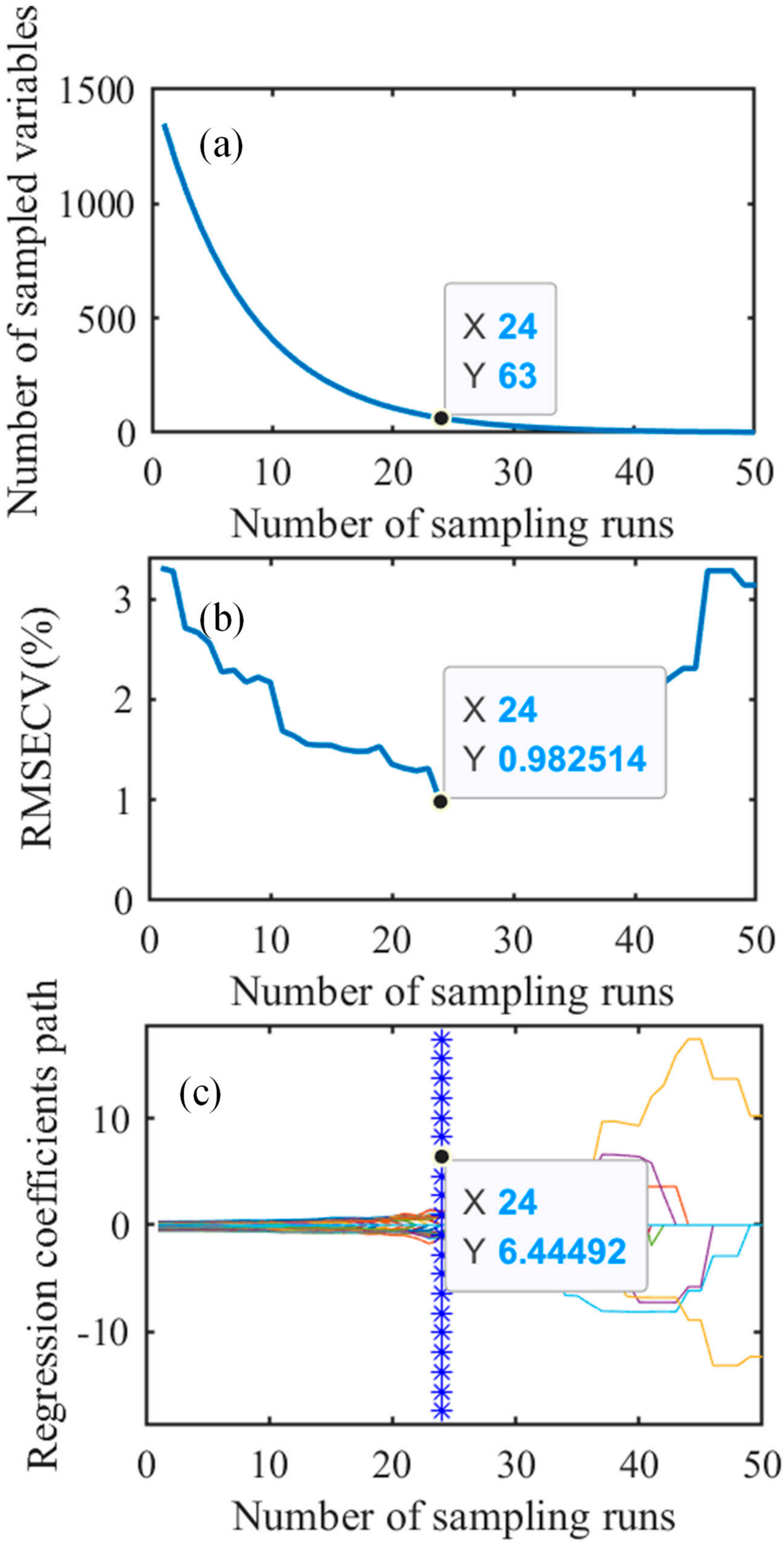

Figure 5. Results of CARS feature extraction for super-θ_f samples. (**a**) Variable change trend, (**b**) Cross validation, (**c**) Trend of regression coefficient. The lines in different color showed the change trend of the regression coefficient of each wavelength with number of sampling runs.

As for the sub-θ_f sample data, when the number of samples was less than 22 (corresponding to the location of "*" in Figure 4c), the RMSECV of the model continuously decreased as the number of samples increased (Figure 4b). The RMSECV reached a minimum value of 1.082 when the number of samplings was 22, indicating that the wavelength variables irrelevant to the SWC had been eliminated at this point. When the number of samples continued to increase, the number of spectra slowly decreased, and the regression coefficient path changed dramatically. This indicates that the redundant variables in the spectra had all been removed before "*"; the wavelength variables removed by the model were correlated with the SWC after "*", and the loss of relevant information caused the RMSECV of the model to gradually increase.

According to the minimum RMSECV in Figure 4b, the optimal subset of feature variables corresponding to the "*" sign was extracted. At this point, the number of sampling times was 22, and the subset corresponding to 83 feature variables was mainly distributed over the ranges 366–370, 642–657, 943–967, 1157–1163, 1214–1221, and 1453–1797 nm (Figure 6).

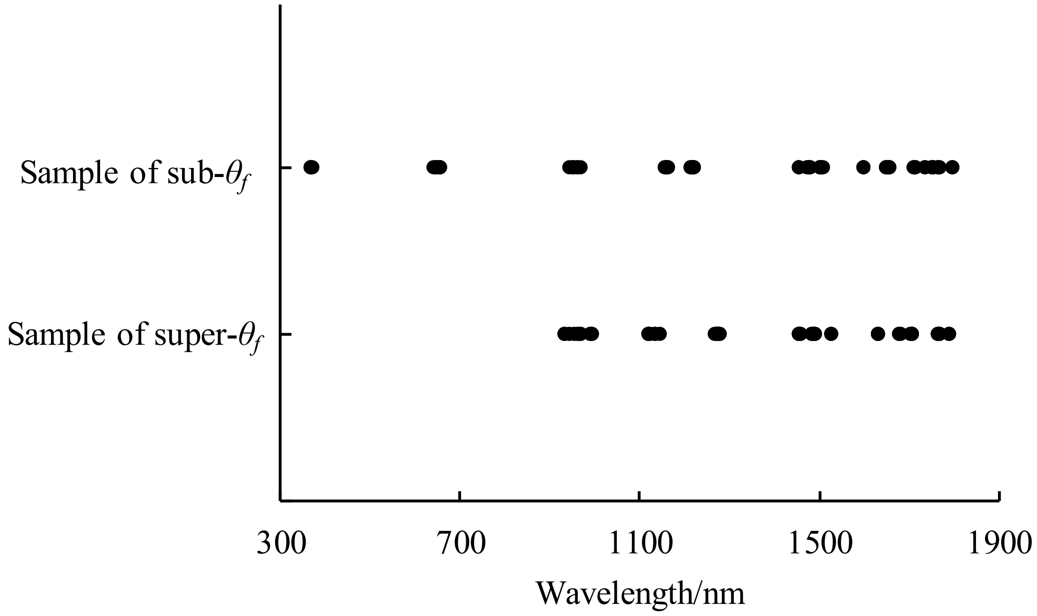

Figure 6. Wavelength distribution of CARS feature extraction.

Similarly, in the super-θ_f sample data, the RMSECV value of the fivefold cross-validation PLSR model was minimized when the number of samples was 24 (at the location corresponding to the "*" sign in Figure 5c), demonstrating that the model worked optimally at this time (Figure 5b). Combined with the minimum RMSECV value in Figure 5b, the optimal subset of spectral variables corresponding to the "*" sign was selected. At this point, the number of sampling times was 24, and the number of feature variables corresponding to the subset was 63, mainly distributed over the ranges 932–993, 1118–1146, 1269–1278, and 1453–1791 nm (Figure 6).

The Rfrog method was used for the secondary screening of the feature extraction results by CARS. The data were processed in a centralized manner. The number of simulation iterations and the initial model variables of Rfrog were set to 10,000 and 6, respectively. The results of each wavelength selection probability are shown in Figure 7a,c.

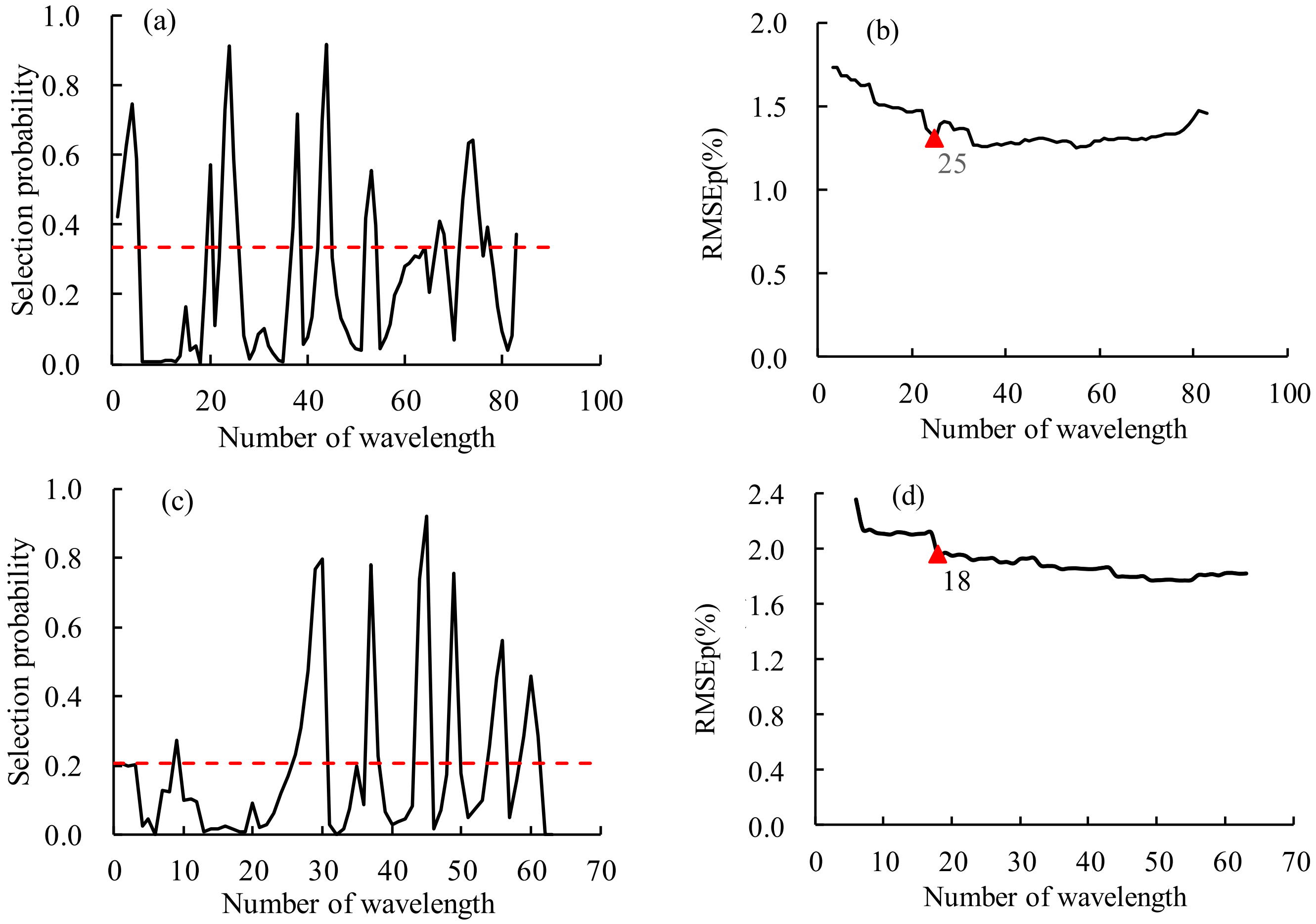

Figure 7. Rfrog run results. (**a**) Wavelength selection probability of sub-θ_f samples, (**b**) determination of threshold of wavelength for the sub-θ_f samples, (**c**) wavelength selection probability of super-θ_f samples, (**d**) determination of threshold of wavelength for the super-θ_f samples.

To extract the feature variables, all the wavelengths were arranged in descending order according to their selection probabilities; the wavelengths were then added one by one to the PLSR model in turn, based on the criterion that a larger selection probability indicated a more important wavelength. The curve representing the RMSE of prediction dataset (RMSEp) of PLSR with the number of selected wavelengths was established (Figure 7b,d).

Through the number of feature wavelengths and the RMSEp value of PLSR, the number of feature variables was determined. For the sub-θ_f samples, the first 25 wavelengths were selected as feature variables, and the RMSEp reached the minimum value of 1.31 (Figure 7b, red triangle position) and corresponded to a wavelength selection probability threshold of 0.33 (Figure 7a, red dashed-line position). The feature wavelengths were mainly distributed over 366–370 nm, 1159–1161 nm, 1454–1479 nm, and 1598–1797 nm (Figure 8). For the super-θ_f samples, the first 18 wavelengths were selected as feature variables, and the RMSEp reached the minimum value of 1.96 (Figure 7d, red triangle position) and corresponded to a wavelength selection probability threshold of 0.21 (Figure 7c, red dashed-line position). The feature wavelengths were mainly distributed over 932 nm, 990 nm, 1274–1278 nm, 1484 nm, 1485 nm, 1525 nm, 1526 nm, and 1678–1768 nm (Figure 8).

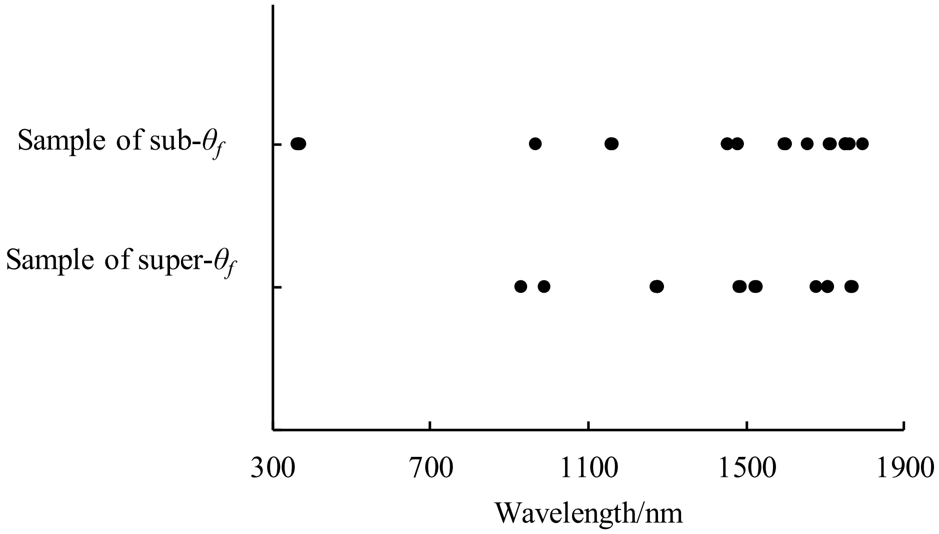

Figure 8. CARS–Rfrog secondary feature extraction wavelength distribution.

3.4. Model Establishment and Evaluation

The feature variables extracted by the CARS–Rfrog method in Section 3.3 were used as independent variables. The ELM, BPANN, and SVM methods were utilized to establish regression models. In the sub-θ_f sample data, all the machine learning models had high accuracy, with R^2c and R^2p > 0.9, RMSEp < 1.6%, and RRMSE < 10% (Table 4). Specifically, the BPANN model had an R^2c of 0.953, R^2p of 0.941, and RRMSE of 6.685%, which was the best model among the three machine learning models. In the super-θ_f sample data (Table 5), the accuracy of all the machine learning models was significantly lower than that of the sub-θ_f sample, and the R^2c and R^2p of all three models were lower than 0.8. The RRMSE of all three models was less than 10%, which could indicate a greater ability to predict the SWC, with the BPANN model having the best results with an R^2c, R^2p, and RRMSE of 0.785, 0.764, and 4.205%.

Table 4. The modeling results of the SWC of sub-θ_f samples according to different models.

Modeling	Calibration Dataset			Prediction Dataset			Parameters
	R^2c	RMSEc/%	RRMSE/%	R^2p	RMSEp/%	RRMSE/%	
ELM	0.931	1.213	5.087	0.924	1.552	6.610	Number of HLNs: 9
BPANN	0.953	1.007	4.225	0.941	1.570	6.685	Number of HLNs: 4
SVM	0.942	1.117	4.688	0.934	1.370	5.837	C = 32, g = 0.0625

Note: HLNs represent the hidden-layer nodes. The same applies subsequently.

Table 5. The modeling results of the SWC of super-θ_f samples according to different models.

Modeling	Calibration Dataset			Prediction Dataset			Parameters
	R^2c	RMSEc/%	RRMSE/%	R^2p	RMSEp/%	RRMSE/%	
ELM	0.751	1.790	4.943	0.748	1.830	5.202	Number of HLNs: 9
BPANN	0.785	1.724	4.759	0.764	1.479	4.205	Number of HLNs: 4
SVM	0.684	2.149	5.933	0.556	1.930	5.485	C = 22.630, g = 0.125

The prediction of SWC by BPANN was optimal for both the sub-θ_f samples and the super-θ_f samples. Considering the evaluation indicators (R^2, RMSE, and RRMSE), the prediction accuracy of the models was BPANN > SVM > ELM for the sub-θ_f samples and BPANN > SVM > ELM for the super-θ_f samples. Figures 9 and 10 show the predicted and measured values of the sub-θ_f samples and super-θ_f samples. The closer the sample points were to the 1:1 line, the better the model prediction ability was. In the sub-θ_f samples for the prediction dataset, the slope of the ELM, BPANN, and SVM fitting line was 1.121, 1.224, and 1.093, respectively. Therefore, the model did not significantly overestimate or underestimate. For the scatter plot of the super-θ_f samples, the points were all more scattered (with a larger RMSE), which also indicated that the prediction accuracy of SWC for the super-θ_f samples needed to be improved.

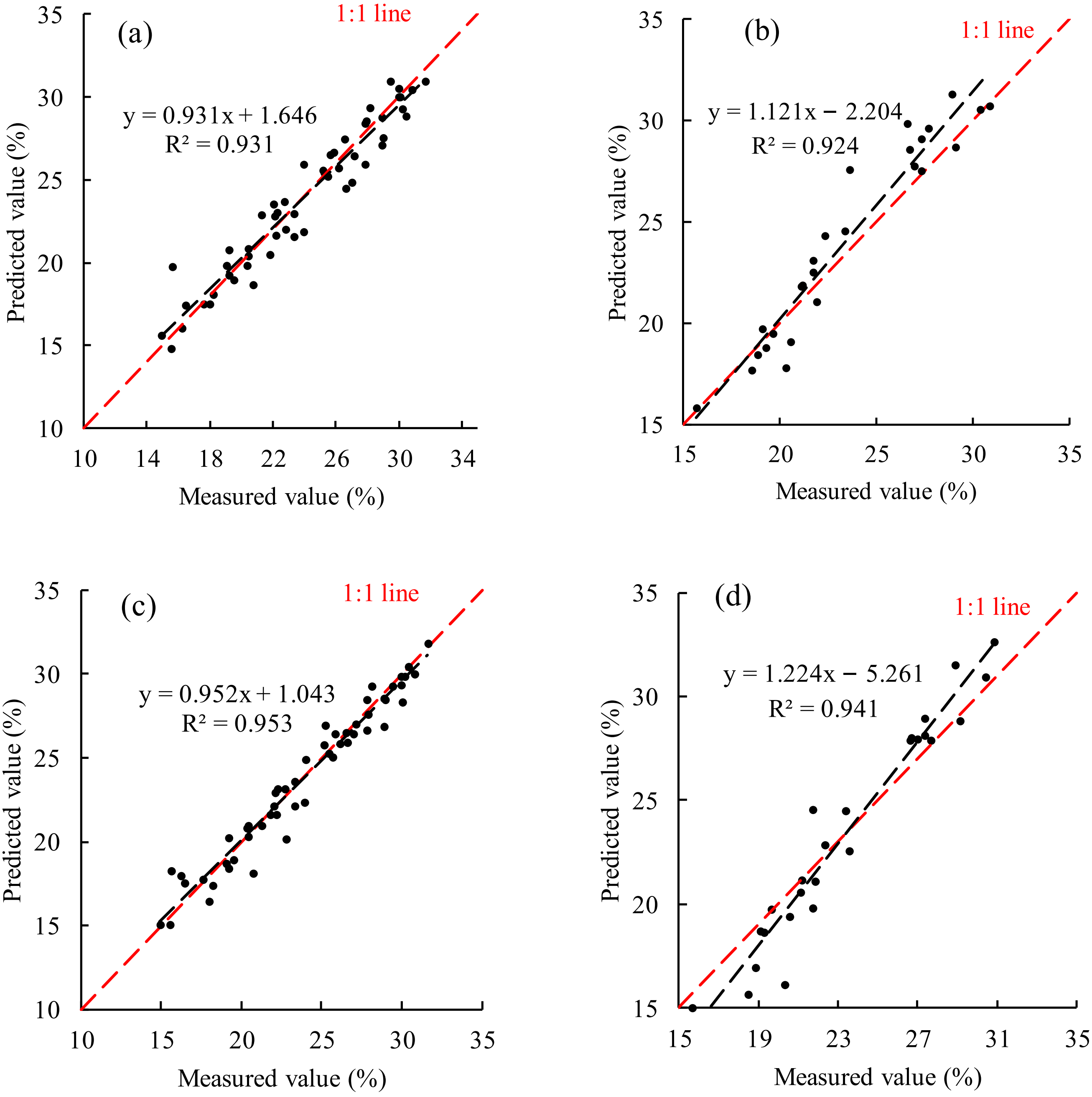

Figure 9. *Cont.*

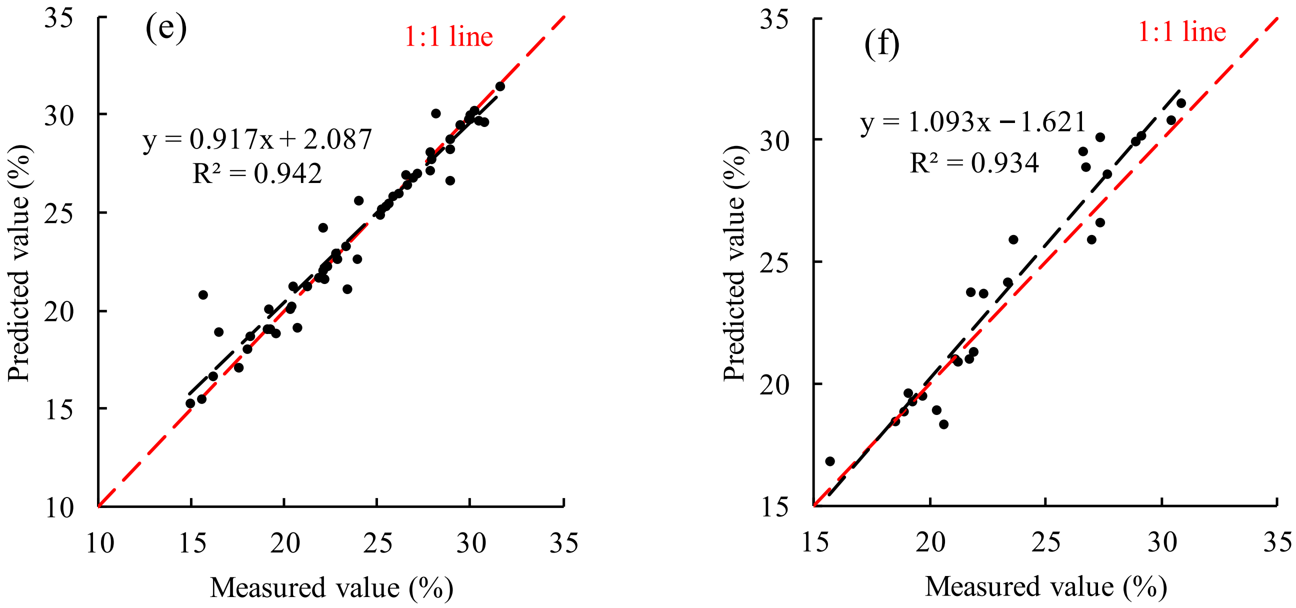

Figure 9. The results of sub-θ_f samples according to different models. (**a**) Calibration dataset, ELM; (**b**) prediction dataset, ELM; (**c**) calibration dataset, BPANN; (**d**) prediction dataset, BPANN; (**e**) calibration dataset, SVM; (**f**) prediction dataset, SVM.

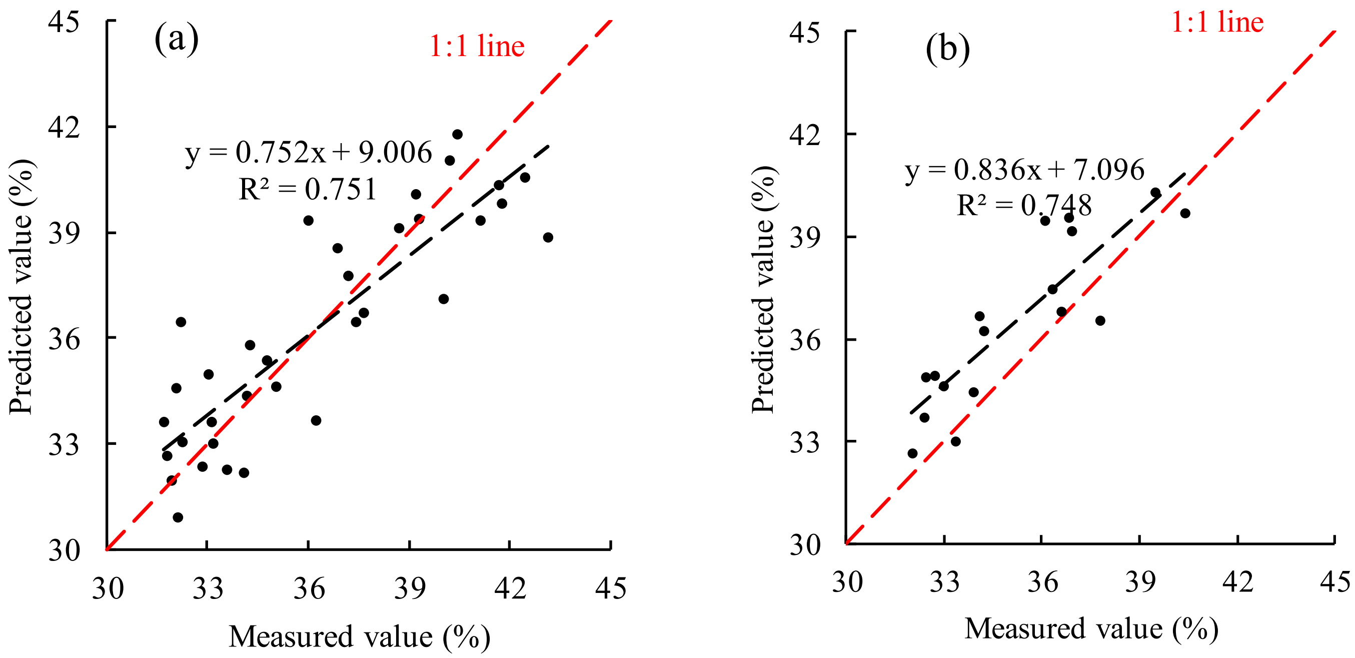

Figure 10. *Cont.*

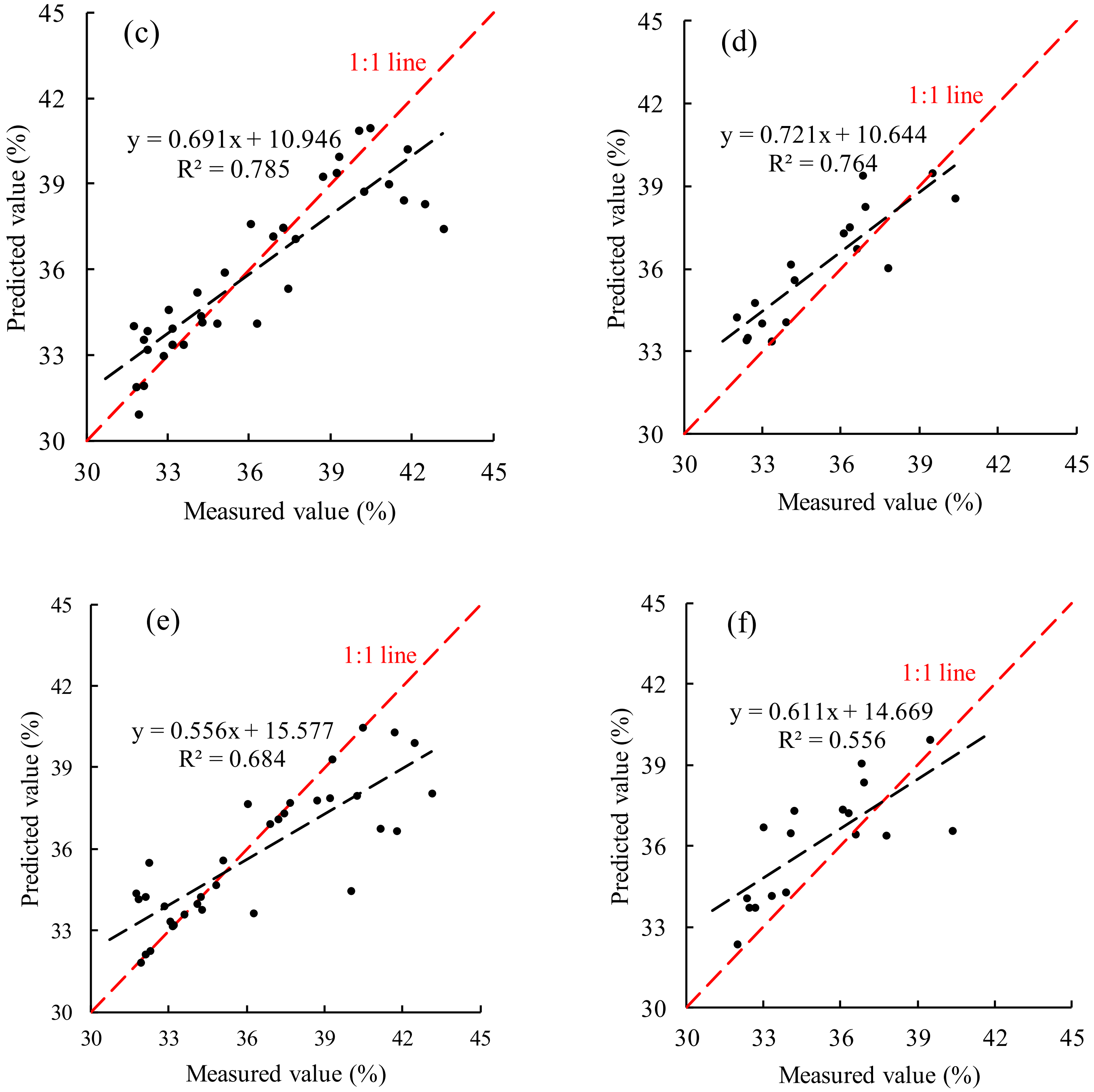

Figure 10. The results of super-θ_f samples according to different models. (**a**) Calibration dataset, ELM; (**b**) prediction dataset, ELM; (**c**) calibration dataset, BPANN; (**d**) prediction dataset, BPANN; (**e**) calibration dataset, SVM; (**f**) prediction dataset, SVM.

4. Discussion

Initially, it was believed that the soil spectral reflectance declined with increasing SWC due to the absorption effect of water on the spectrum [12,35]. However, subsequent studies [14] have shown that when the SWC exceeds θ_f, a water film forms on the surface of soil particles, resulting in specular reflection and causing the soil spectral reflectance to increase with increasing SWC. Previous studies tended to control the measured sample SWC to remain below θ_f when establishing a prediction model of SWC based on hyperspectral reflectance or possibly did not distinguish whether the SWC was higher or lower than the θ_f. Nonetheless, considering that the soil reflectance changing process decreased and then increased with θ_f as the threshold and the full-spectrum modeling of samples by PLSR in this paper, it was found that the model performed poorly in the super-θ_f samples, with the accuracy being much lower than that of the sub-θ_f sample part of the model. Therefore, this paper modeled samples with water content below θ_f and above θ_f separately to improve the accuracy of the SWC prediction model based on the hyperspectral data.

The feature variables extracted by CARS–Rfrog in this paper were mainly concentrated in the near-infrared (NIR) region, with almost none in the visible band (except for the 366–370 nm range). This is consistent with the findings of previous studies [36,37]. The NIR spectrum is generated due to the vibrational energy level jumps and rotational energy level jumps in molecules. When the vibration and rotation of a molecule jump from the ground state or low-energy level to a higher-energy state, they absorb a certain amount of infrared energy from the external incident electromagnetic radiation. In the mid-infrared region, fundamental frequency absorption occurs, while in the NIR region, combined frequency and doubled frequency absorption occur. There are three fundamental frequencies of water molecules in the near-infrared band [38]. Therefore, the NIR region of the band can reflect changes in soil moisture.

In this study, secondary extraction (CARS–Rfrog) was applied to extract variables of hyperspectral reflectance. Firstly, CARS was used for the initial screening of the feature variables. However, the CARS extraction results were random [20], and the feature variables extracted by only one method are numerous, making the model too complex to model.

Therefore, the feature variables extracted by CARS were subjected to a second extraction using the Rfrog method to obtain the variables with the least redundant information. By retaining bands with high correlation and downscaling and re-extracting the feature variables, we simplify the model. The Rfrog method applied the partial least squares linear discriminant analysis to construct the classifier and combined the strong ability of synergy interval partial least squares (SiPLS) to handle highly correlated data [32].

The Rfrog method combines the ideas of the memetic algorithm and the particle swarm optimization algorithm, so it has the characteristics of survival of the fittest and random search. It also takes advantage of CARS to simplify the complexity of wavelength selection. Through secondary feature extraction, the number of bands in the modeled data was greatly reduced. Rfrog reduced the number of variables of the super-θ_f and sub-θ_f samples extracted using CARS from 83 and 63 to 25 and 18, respectively. CARS–Rfrog minimized the redundant information and achieved the effect of data dimensionality reduction.

Among the three machine learning models, the BPANN model had the highest accuracy (Tables 3 and 4), which could deliver a better prediction of SWC. This might mean that the SVM is more suitable for fewer-sample modeling, while the ELM tends to have low and unstable prediction accuracy when dealing with the quantitative analysis of complex samples [39]. However, the ELM is fast learning and has a strong generalization ability, so it is frequently used in scenarios that require real-time computing. The BPANN is expressive and simple, and the theory also demonstrated that a three-layer neural network could approximate a nonlinear continuous function with arbitrary accuracy, which makes it possible to solve complex nonlinear problems with internal mechanisms. However, its generalization ability is slightly inferior, and it easily falls into locally optimal solutions [40], and subsequent studies have attempted to apply optimization algorithms to optimize BPANN and obtain a model with better performance. Therefore, CARS–Rfrog–BPANN is recommended as a prediction model for the SWC of red soil.

5. Conclusions

In this study, soil samples were prepared in the laboratory, and the hyperspectral reflectance was acquired outdoors. The samples were divided into two parts (sub-θ_f and super-θ_f) with θ_f as a threshold to obtain a more accurate SWC prediction model. The outliers were detected using MCCV; the spectral feature variables were extracted using a secondary extraction method (CARS–Rfrog), and the prediction model of SWC was established using the machine learning method. We draw the following conclusions:

(1) The poor performance of the model in the fraction of water content above the θ_f when the model was built with full-spectrum PLSR, indicated that using the same model for the simultaneous inversion of SWC under both conditions of water content above or below θ_f led to poor inversion accuracy of samples above the θ_f. (2) By combining CARS and Rfrog for the extraction of the feature variables of soil reflectance, the feature wavelengths of the sub-θ_f and super-θ_f samples extracted by CARS–Rfrog were 25 and 18, and they were widely distributed in the NIR range, which is a significant reduction in comparison to the full spectrum. (3) Among the machine learning methods, the BPANN achieved optimal prediction results, the R^2p, RMSE$_P$, and RRMSE of the sub-θ_f samples were 0.941, 1.570%, and 6.685%, respectively, and the R^2p, RMSEp, and RRMSE of the super-θ_f samples were 0.764, 1.479%, and 4.205%, respectively.

Author Contributions: Conceptualization, F.L. and S.C.; methodology, F.L. and S.C.; software, F.L. and S.C.; validation, K.P., S.Z. and S.C.; formal analysis, Y.T., S.T. and Q.W.; investigation, F.L., K.P. and S.Z.; data curation, F.L., K.P. and S.Z.; writing—original draft preparation, F.L. and S.C.; writing—review and editing, Y.T., S.T. and Q.W.; visualization, F.L. and S.C.; supervision, Y.T., S.T. and Q.W.; funding acquisition, S.C. and S.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number [52209056, 52009039]; the Yunnan Fundamental Research Projects, grant number. [202301AU070190, 202301AU070095]; the Yunnan Science and Technology Talent and Platform Program, grant number [202305AM070006]; and the Scientific Research Fund Project of the Yunnan Provincial Department of Education, grant number [2022J0061].

Data Availability Statement: The data used to support the findings of this study are included within the article.

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

References

1. Li, F.; Peng, X.F.; Chen, X.W.; Liu, M.L.; Xu, L.W. Analysis of Key Issues on GNSS-R Soil Moisture Retrieval Based on Different Antenna Patterns. *Sensors* **2018**, *18*, 2498. [CrossRef]
2. Brocca, L.; Ciabatta, L.; Massari, C.; Camici, S.; Tarpanelli, A. Soil Moisture for Hydrological Applications: Open Questions and New Opportunities. *Water* **2017**, *9*, 140. [CrossRef]
3. Eroglu, O.; Kurum, M.; Boyd, D.; Gurbuz, A.C. High Spatio-Temporal Resolution CYGNSS Soil Moisture Estimates Using Artificial Neural Networks. *Remote Sens.* **2019**, *11*, 2272. [CrossRef]
4. Bittelli, M. Measuring Soil Water Content: A Review. *Horttechnology* **2011**, *21*, 293–300. [CrossRef]
5. Dobriyal, P.; Qureshi, A.; Badola, R.; Hussain, S.A. A review of the methods available for estimating soil moisture and its implications for water resource management. *J. Hydrol.* **2012**, *458*, 110–117. [CrossRef]
6. Abdelmoneim, A.A.; Khadra, R.; Derardja, B.; Dragonetti, G. Internet of Things (IoT) for Soil Moisture Tensiometer Automation. *Micromachines* **2023**, *14*, 263. [CrossRef]
7. Steele-Dunne, S.C.; Rutten, M.M.; Krzeminska, D.M.; Hausner, M.; Tyler, S.W.; Selker, J.; Bogaard, T.A.; de Giesen, N.C.V. Feasibility of soil moisture estimation using passive distributed temperature sensing. *Water Resour. Res.* **2010**, *46*, W03534. [CrossRef]
8. Sayde, C.; Buelga, J.B.; Rodriguez-Sinobas, L.; El Khoury, L.; English, M.; van de Giesen, N.; Selker, J.S. Mapping variability of soil water content and flux across 1–1000 m scales using the Actively Heated Fiber Optic method. *Water Resour. Res.* **2014**, *50*, 7302–7317. [CrossRef]
9. Hu, Y.; Li, M.; Ren, H.; Si, B.C. Measurement of soil water content using distributed temperature sensor with heated fiber optics. *Trans. Chin. Soc. Agric. Eng.* **2019**, *35*, 42–49. (In Chinese)
10. Rijal, S.; Zhang, X.D.; Jia, X.H. Estimating Surface Soil Water Content in the Red River Valley of the North using Landsat 5 TM Data. *Soil Sci. Soc. Am. J.* **2013**, *77*, 1133–1143. [CrossRef]
11. Yang, X.G.; Yu, Y. Remote sensing inversion of soil moisture based on laboratory spectral reflectance data. *Trans. Chin. Soc. Agric. Eng.* **2017**, *33*, 195–199. (In Chinese)
12. Bowers, S.A.; Hanks, R.J. Reflection of radiant energy from soils. *Soil Sci.* **1965**, *100*, 130–138. [CrossRef]
13. Lobell, D.B.; Asner, G.P. Moisture effects on soil reflectance. *Soil Sci. Soc. Am. J.* **2002**, *66*, 722–727. [CrossRef]
14. Muller, E.; Décamps, H. Modeling soil moisture-reflectance. *Remote Sens. Environ.* **2001**, *76*, 173–180. [CrossRef]
15. Neema, D.L.; Shah, A.; Patel, A.N. A statistical optical model for light reflection and penetration through sand. *Int. J. Remote Sens.* **1987**, *8*, 1209–1217. [CrossRef]

16. Liu, W.D.; Baret, F.; Gu, X.F.; Tong, Q.X.; Zheng, L.F.; Zhang, B. Relating soil surface moisture to reflectance. *Remote Sens. Environ.* **2002**, *81*, 238–246.
17. Sun, Y.J.; Zheng, X.P.; Qin, Q.M.; Meng, Q.Y.; Gao, Z.l.; Ren, H.Z.; Wu, L. Modeling Soil Spectral Reflectance with Different Mass Moisture Content. *Spectrosc. Spectr. Anal.* **2015**, *35*, 2236–2240. (In Chinese)
18. Bailey-Serres, J.; Lee, S.C.; Brinton, E. Waterproofing Crops: Effective Flooding Survival Strategies. *Plant Physiol.* **2012**, *160*, 1698–1709. [CrossRef]
19. Yun, Y.H.; Li, H.D.; Deng, B.C.; Cao, D.S. An overview of variable selection methods in multivariate analysis of near-infrared spectra. *Trac-Trends Anal. Chem.* **2019**, *113*, 102–115. [CrossRef]
20. Xu, L.J.; Chen, M.; Wang, Y.C.; Chen, X.Y.; Lei, X.L. Study on Non-Destructive Detection Method of Kiwifruit Sugar Content Based on Hyperspectral Imaging Technology. *Spectrosc. Spectr. Anal.* **2021**, *41*, 2188–2195. (In Chinese)
21. Ye, S.F.; Wang, D.; Min, S.G. Successive projections algorithm combined with uninformative variable elimination for spectral variable selection. *Chemom. Intell. Lab. Syst.* **2008**, *91*, 194–199. [CrossRef]
22. Xu, Y.C.; Wang, X.Y.; Yin, X.; Hu, Z.X.; Yue, R.C. Visualization Spatial Assessment of Potato Dry Matter. *Trans. Chin. Soc. Agric. Mach.* **2018**, *49*, 339–344. (In Chinese)
23. Sun, J.Y.; Li, M.Z.; Tang, L.H.; Zheng, L.H. Spectral Characteristics and Their Correlation with Soil Parameters of Black Soil in Northeast China. *Spectrosc. Spectr. Anal.* **2007**, *27*, 1502–1505. (In Chinese)
24. Tian, J.; Philpot, W.D. Relationship between surface soil water content, evaporation rate, and water absorption band depths in SWIR reflectance spectra. *Remote Sens. Environ.* **2015**, *169*, 280–289. [CrossRef]
25. Oltra-Carrio, R.; Baup, F.; Fabre, S.; Fieuzal, R.; Briottet, X. Improvement of Soil Moisture Retrieval from Hyperspectral VNIR-SWIR Data Using Clay Content Information: From Laboratory to Field Experiments. *Remote Sens.* **2015**, *7*, 3184–3205.
26. Sekertekin, A.; Marangoz, A.M.; Abdikan, S. ALOS-2 and Sentinel-1 SAR data sensitivity analysis to surface soil moisture over bare and vegetated agricultural fields. *Comput. Electron. Agric.* **2020**, *171*, 105303. [CrossRef]
27. Shang, T.H.; Jia, P.P.; Sun, Y.; Zhang, J.H. Spectral Characteristics of Soil Moisture in Salinized Soil and Model Fitting Accuracy in Northern Yinchua City, Ningxia Hui Autonomous Region. *Bull. Soil Water Conserv.* **2020**, *40*, 183–189. (In Chinese)
28. Diao, W.Y.; Liu, G.; Hu, K.L. Estimation of Soil Water Content Based on Hyperspetral Features and the ANN Model. *Spectrosc. Spectr. Anal.* **2017**, *37*, 841–846. (In Chinese)
29. Cao, D.S.; Liang, Y.Z.; Xu, Q.S.; Li, H.D.; Chen, X. A new strategy of outlier detection for QSAR/QSPR. *J. Comput. Chem.* **2010**, *31*, 592–602. [CrossRef]
30. Galvao, R.K.H.; Araujo, M.C.U.; Jose, G.E.; Pontes, M.J.C.; Silva, E.C.; Saldanha, T.C.B. A method for calibration and validation subset partitioning. *Talanta* **2005**, *67*, 736–740. [CrossRef]
31. Li, H.D.; Liang, Y.Z.; Xu, Q.S.; Cao, D.S. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Anal. Chim. Acta* **2009**, *648*, 77–84. [CrossRef]
32. Li, H.D.; Xu, Q.S.; Liang, Y.Z. Random frog: An efficient reversible jump Markov Chain Monte Carlo-like approach for variable selection with applications to gene selection and disease classification. *Anal. Chim. Acta* **2012**, *740*, 20–26. [CrossRef]
33. Huang, G.B.; Zhu, Q.Y.; Siew, C.K. Extreme learning machine: Theory and applications. *Neurocomputing* **2006**, *70*, 489–501. [CrossRef]
34. Vapnik, V. *Estimation of Dependences Based on Empirical Data*; Springer: New York, NY, USA, 2006.
35. Zhu, Y.H.; Deng, R.D.; Lu, Y.F.; Chen, M.Z. Varying characteristics of spectral reflectivity in different humidities of yellow-brown earth and its significance in renote sensing. *Acta Pedol. Sin.* **1984**, *21*, 194–202. (In Chinese)
36. Wang, C.Z.; Wang, J.H.; Wang, J.D.; Zhao, C.J.; Liu, L.J.; Wang, P.X.; Jing, J.J. The Choice of Best Detecting Band for Hyperspectral Remote Sensing on Surface Water Content of Bare Soil. *Remote Sens. Inf.* **2003**, *4*, 33–36. (In Chinese)
37. Yao, Y.M.; Wei, N.; Tang, P.Q.; Li, Z.B.; Yu, Q.Y.; Xu, X.G.; Chen, Y.Q.; He, Y.B. Hyper-spectral characteristics and modeling of black soil moisture content. *Trans. Chin. Soc. Agric. Eng.* **2011**, *27*, 95–100.
38. Shi, Z. *Principles and Methods of Ground-Based Hyperspectral Remote Sensing of Soils*; Science Press: Beijing, China, 2014. (In Chinese)
39. Bian, X.H.; Zhang, C.X.; Tan, X.Y.; Dymek, M.; Guo, Y.G.; Lin, L.G.; Cheng, B.W.; Hu, X.Y. A boosting extreme learning machine for near-infrared spectral quantitative analysis of diesel fuel and edible blend oil samples. *Anal. Methods* **2017**, *9*, 2983–2989. [CrossRef]
40. Liu, W.; Liu, S.; Bai, R.C.; Zhou, X.; Zhou, D.N. Research of Mutual Learning Neural Network Training Method. *Chin. J. Comput.* **2017**, *40*, 1291–1308. (In Chinese)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.