Article

Estimation of Soil Organic Matter Based on Spectral Indices Combined with Water Removal Algorithm

Jiawei Xu 1,2, Yuteng Liu 1,2, Changxiang Yan 1,3 and Jing Yuan 1,*

1 Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China; xujiawei211@mails.ucas.ac.cn (J.X.); liuyuteng22@mails.ucas.ac.cn (Y.L.); yancx@ciomp.ac.cn (C.Y.)
2 University of Chinese Academy of Sciences, Beijing 100049, China
3 Center of Materials Science and Optoelectrics Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: yuanjing@ciomp.ac.cn; Tel.: +86-131-0440-2340

Abstract: Soil moisture strongly interferes with the spectra of soil organic matter (SOM) in the near-infrared region, which reduces the correlation between organic matter and spectra and decreases accuracy in the prediction of SOM. In this study, we explored the feasibility of two types of spectral indices, two- and three-band mixed (SI) and three-band spectral indices (SI3), and two water removal algorithms, direct standardization (DS) and external parameter orthogonalization (EPO), to estimate SOM in wet soils using a total of 192 soil samples at six water content gradients. The estimation accuracies of spectral indices combined with water removal algorithms were better than those of full spectral data combined with water removal algorithms: the prediction accuracies of SI-EPO ($R^2 = 0.735$, RMSEP = 3.4102 g/kg) were higher than those of EPO ($R^2 = 0.63$, RMSEP = 4.1021 g/kg), and those of SI-DS ($R^2 = 0.70$, RMSEP = 3.7085 g/kg) were higher than those of DS ($R^2 = 0.61$, RMSEP = 4.2806 g/kg); SI3-EPO ($R^2 = 0.752$, RMSEP = 3.1344 g/kg) was better than SI-EPO; both EPO and DS effectively mitigated the influence of soil moisture, with EPO demonstrating superior performance in small-sample prediction scenarios. This study introduces a novel approach to counteract the impact of soil moisture on SOM estimation.

Keywords: soil spectroscopy; soil organic matter; soil moisture; spectral index; external parameter orthogonalization; direct standardization

1. Introduction

Black soil, abundant in organic matter, serves as a vital substrate for cultivating staple food crops, alongside other crops and livestock. However, this soil type is highly susceptible to degradation through overuse, erosion, and contamination. Consequently, safeguarding black soil is imperative, not only for future agricultural productivity and food security but also for preserving soil health and promoting sustainable soil management practices. Organic matter, a key component of soil fertility, serves as a significant reservoir of nutrients, including nitrogen, phosphorus, and potassium, which are essential for plant growth. Its decomposition enriches agricultural soils, thereby enhancing crop yields and contributing significantly to soil ecosystem functioning, agricultural productivity, and environmental conservation [1]. Hyperspectral technology plays an important role in air quality detection, ultra-short laser pulse characterization, spectral imaging, the detection of soil properties, and other fields [2]. Field visible-near-infrared (vis-NIR) soil spectroscopy represents a promising alternative to traditional soil organic matter assessment methods. By leveraging remote sensing satellites or ground sensors, vis-NIR spectroscopy streamlines data collection, offering advantages such as efficiency, broad coverage, environmental friendliness, and cost-effectiveness. In contrast...
to conventional field data collection methods, which demand significant human, material, and financial resources and entail long processing times and limited coverage, vis-NIR spectroscopy presents a convenient, rapid, and affordable solution for obtaining soil property data, including organic matter content [3,4].

Various studies by domestic and international scholars have explored modeling and predicting soil properties using visible near-infrared (VNIR) reflectance hyperspectral techniques (350–2500 nm) [5]. Machine learning algorithms, known for their nonlinear capabilities and robust data mining prowess, have been instrumental in constructing regression models for estimating soil organic matter (SOM) content based on hyperspectral remote sensing data [6–8]. For instance, Yu et al. determined that a partial least squares regression (PLSR) model, combined with continuous uniform removal transformation, yielded the optimal prediction model for SOM content in the Jianghan Plain [9]. Additionally, Chang introduced a dynamic adaptive inertia-weighted particle swarm optimization (DPSO) algorithm to fine-tune the parameters of an artificial neural network (ANN), resulting in enhanced SOM prediction accuracy [10]. Ou addressed overfitting issues by employing a semi-supervised deep neural network regression (Semi-DNNR) model, which incorporated various parametric inversion features [11]. Zhang utilized a genetic algorithm (GA) for band selection, leading to improved predictive model calibration with partial least squares regression (PLSR) [12]. However, it is important to note that these methods typically utilized dried soil samples for modeling and prediction purposes [13–15].

Minor fluctuations in soil moisture content can significantly impact the way microwaves are reflected and scattered at the soil surface [16]. Soil moisture serves as a significant factor in spectral analysis and the determination of soil organic matter (SOM) content in the near-infrared (NIR) region, disrupting the correlation between SOM and spectral data. Among soil components, soil moisture content (SMC) exerts the most substantial influence on the range of soil reflectance in the visible near-infrared (VNIR) region [17–19]. Previous research has delved into the effects of SMC on reflectance by rehydrating soil samples, demonstrating that SMC can impact the accuracy of SOM estimation [20,21]. SMC can interfere with SOM prediction by obscuring key soil spectral absorption features and distorting the overall spectral profile [22].

To remove the water effect from the wet spectrum, researchers have used various methods. Yang et al. used an external parameter orthogonalization (EPO) algorithm to remove the effect of soil water content and improved the accuracy of SOM content estimation [23]. Direct standardization (DS) and piecewise direct standardization (PDS) have also been used to remove the effect of water [24]. EPO can eliminate external influences orthogonal to the target characteristics, thus allowing for the removal of known external factors from spectral features. EPO requires a dataset consisting of the same samples from two different measurement conditions. Similarly, DS requires the transmission of sample spectra to the desired reference space through a transfer set [25].

Hong used the Normalized Soil Moisture Index (NSMI) to group soil water content and estimated SOM content and concluded that NSMI can be used as a measure of soil water content [26]. Chen et al. used Singular Value Decomposition (SVD) combined with correlation analysis to screen soil moisture characteristic spectra, constructed correction coefficients for removing moisture factors to form correction spectra for wet soil spectra, and found that the estimation accuracy was improved after modeling analysis [27].

Most of the above studies used spectral data in full spectral bands for the predictive modeling of soil properties. Soil hyperspectral curves contain many bands, and there are different degrees of correlation between bands. There is a certain amount of information redundancy. Extracting the soil spectral feature index can reduce the input variables for the spectral modeling of soil properties and improve the modeling efficiency.

The aims of this study were as follows:
To analyze the effect of moisture content on soil spectral reflectance and compare the ability of external parameter orthogonalization (EPO) and direct standardization (DS) to eliminate the effect of moisture in VNIR spectra.

To create a two-band spectral index and a three-band spectral index, investigate their correlation with soil organic matter content, and identify the most sensitive bands.

To construct an improved algorithm for combining SOM spectral indices with EPO and DS to overcome and eliminate the effect of soil moisture on the accuracy of SOM spectral monitoring in order to improve the accuracy of SOM detection.

2. Materials and Methods

2.1. Experimental Materials

2.1.1. Study Area and Soil Data Collection

Soil samples were collected from Gongzhuling City, Jilin Province, China. Gongzhuling City is located in the central and western part of Jilin Province, on the right bank of the middle course of the Dongliao River, in the middle of the Songliao Plain, between 124°02′ to 125°18′E and 43°11′ to 44°09′N (Figure 1). The geological structure is covered by alluvial layers throughout the season, most of which are composed of black clay. A few areas contain sandy clay, which is a good research object for black soil.

The experimental field was divided into 32 experimental plots based on organic matter concentration. Within each plot, we employed a five-point sampling technique. Soil samples were divided into two portions: one for measuring soil organic matter content, and the other for collecting spectral data. The soil samples were air-dried, ground, and sieved through a 2 mm mesh. Organic matter content was determined using the Agricultural Soil Inventory (ASI) method. The soil samples exhibited a range of organic matter contents, with a maximum of 38.60 g/kg, a minimum of 17.98 g/kg, and an average of 27.75 g/kg.

Figure 1. Soil collection locations in the study area.
2.1.2. Soil Moisture Content Setting

Six gradients were set for soil moisture content, namely 12%, 14%, 16%, 18%, 20%, and 22%. In Formula (1), \(W_2\) is the weight of wet soil, \(W_1\) is the weight of dry soil, and \(W_3\) is the weight of the plastic box containing the soil. We added a certain amount of water according to the formula, sealed the container, and let it sit for one night to achieve the corresponding moisture content gradient. In this way, 192 soil samples with different organic matter contents and moisture contents were obtained, which were used to establish a soil organic matter hyperspectral prediction model.

\[
\theta_m = \frac{W_2 - W_1}{W_3 - W_1} \times 100\%
\]

(1)

2.1.3. Spectral Measurement

Hyperspectral measurements were conducted in the laboratory using the Field-Spec 3 Hi-Res spectrometer from the American ASD Company (Alpharetta, GA, USA) on 192 wet soil samples and 32 dry soil samples. The spectrometer’s range spans from 350 to 2500 nm, employing a 50 W halogen lamp as its light source, set at a 45° incident angle, positioned 20 cm away from the soil sample. A lens added to the probe narrowed the field of view angle to 1°, with the probe placed 10 cm from the soil sample. Each measurement automatically collected 10 spectral curves, with the arithmetic average utilized as the spectral curve for each soil sample. Standard white plate calibration was performed prior to each measurement.

2.1.4. Data Preprocessing

Since 350–500 nm and 2400–2500 nm are affected by the low light sensitivity of the ASD and low soil reflectivity [28], the signal-to-noise ratio was low, so this part of the band was eliminated during the study, and only the soil reflection in the 500–2400 nm band was analyzed. The Savitzky–Golay smoothing method was used to smooth the spectral data after correcting the breakpoints. The Savitzky–Golay smoothing method was proposed by Savitzky and Golay in 1964 [29]. It is a digital signal processing filter based on polynomial fitting. It performs polynomial fitting within a local area of the signal and uses the points on the fitting curve as the filtered output value. This filtering method has a certain smoothing effect while retaining the overall trend of the signal.

2.2. Experimental Methods

2.2.1. Spectral Index

Spectral indices play a crucial role in various fields, such as remote sensing, environmental science, and agriculture, offering insights into vegetation condition, soil texture, water quality, and more. They are calculated from spectral bands and come in one-, two-, and three-band types. While the first-band spectral index is straightforward but limited in data, second- and third-band indices provide more comprehensive information.

This study employed three common spectral index calculation methods, the difference spectral index (DSI), ratio spectral index (RSI), and normalized difference spectral index (NDSI), as outlined in Table 1.

<table>
<thead>
<tr>
<th>Spectral Index</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difference spectral index</td>
<td>(DSI(\lambda_i, \lambda_j) = R(\lambda_i) - R(\lambda_j))</td>
</tr>
<tr>
<td>Ratio spectral index</td>
<td>(RSI(\lambda_i, \lambda_j) = R(\lambda_i) / R(\lambda_j))</td>
</tr>
</tbody>
</table>
SI2 Normalized difference spectral index

$$NDSI(\lambda_i, \lambda_j) = \frac{R(\lambda_i) - R(\lambda_j)}{R(\lambda_i) + R(\lambda_j)}$$

SI3 Normalized difference spectral index

$$NDSI(\lambda_i, \lambda_j, \lambda_k) = \frac{aR(\lambda_i) + bR(\lambda_j) + cR(\lambda_k)}{dR(\lambda_i) + eR(\lambda_j) + fR(\lambda_k)}$$

$$a, b, c, d, e, f \in [-1, 0, 1]$$

$$R(\lambda_i)$$ and $$R(\lambda_j)$$ in the formulas indicate the reflectance value for any two wavelengths $$\lambda_i$$ and $$\lambda_j$$ in 500–2400 nm.

2.2.2. External Parameter Orthogonalization (EPO)

External parameter orthogonalization (EPO) was first proposed by Roger et al. [30]. The proposed EPO-PLS method corrects the influence of temperature variables on the spectral prediction of sugar content in fruits. Minasny et al. applied EPO to soil organic carbon measurements [31]. The use of EPO can eliminate the influence of soil moisture from the spectrum, thereby improving the calibration and prediction of soil organic carbon content.

The influence of SMC on wet soil spectra can be represented by a matrix as follows:

$$X = XP + XQ$$

In the formula, $$X$$ represents the matrix of wet soil spectra. $$P$$ is the projection matrix that contains useful information about SOM in $$X$$. $$Q$$ is the projection matrix that represents the useless information generated due to the influence of SMC.

The core idea of the EPO is to extract the projection matrix $$P$$ that covers useful spectral information from $$X$$. The specific algorithm process is as follows:

1. Dry soil samples can be represented as $$(n \times m)$$, and the average spectrum of dry soil samples can be calculated as $$(1 \times m)$$.
2. Wet soil samples can be represented as $$(N \times m)$$, and the average spectrum of samples with different SMCs can be calculated as $$(h \times m)$$.
3. Calculate the difference between dry and wet soil samples as $$D$$ $$(h \times m)$$.
4. Perform Principal Component Analysis on DTD to obtain the matrix $$V$$ $$(m \times m)$$.
5. Define the dimension of EPO as $$g$$ and calculate a subset $$Vs$$ $$(m \times g)$$ of the $$V$$ matrix.
6. Calculate $$Q = VsVs^T$$.
7. Finally, calculate the projection matrix $$P$$ using the formula $$P = I - Q$$, where $$I$$ is an identity matrix.

2.2.3. Direct Standardization (DS)

Direct standardization (DS) [32] is a commonly used spectral data preprocessing method that normalizes the spectral differences between different samples, making the spectral data easier to compare and analyze. DS was initially used for transfer calibration between VIS-NIR spectrometers. Ji et al. [33] proposed the use of DS to remove the influence of SMC in soil spectra.

The dry soil spectrum $$X$$ can be represented as follows:

$$X = XwB + \lambda d_\lambda$$

where $$B$$ is the transfer matrix, and $$\lambda d_\lambda$$ is the residual term.

The specific steps of the DS are as follows: Collect a set of soil samples, including samples with different SMCs. Perform spectral measurements on each soil sample to obtain spectral data. Select a specific wavelength range where the spectral signal is primarily influenced by moisture. For the spectral data within this wavelength range, for each soil sample, calculate the average. For each soil sample, subtract the corresponding average value from the spectral data within this wavelength range.
2.2.4. Model Construction Methods and Accuracy Evaluation Indicators

The SOM prediction model adopts the partial least squares (PLS) method. PLS is a statistical method used to build linear regression models and perform data dimensionality reduction. It is mainly used for handling multivariate data and finds widespread applications in tasks such as regression analysis and data prediction.

The 192 samples from the experiment were split into a training set comprising 160 samples and a validation set with 32 samples. The training set samples were used to analyze the correlation between spectral indices and SOM content, build SOM prediction models based on EPO and DS, and improve the SOM prediction model based on the spectral indices. The validation set was solely employed to assess the model's accuracy.

The accuracy of the PLS model was evaluated using the coefficient of determination ($R^2$), bias, and the root mean squared error (RMSE) [12] between the predicted and measured SOM content. A higher $R^2$ and lower RMSE indicate that the model has greater prediction accuracy. The closer the bias is to 0, the higher the accuracy of the model.

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

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}$$

where $y_i$ and $\hat{y}_i$ are the observed values and the predicted values, $\bar{y}$ is the mean of the observed values, and $N$ is the number of samples with $i$ equal to 1, 2, … $N$.

3. Results

3.1. Effect of Soil Moisture Content on Soil Spectral Reflectance

Figure 2a illustrates the average spectral reflectance of soil across seven different soil moisture content (SMC) gradients. Throughout the entire spectral range from 500 to 2400 nm, there is a noticeable decrease in spectral reflectance as the moisture content increases. This decrease is primarily due to the increased refraction of light between soil particles with a higher SMC, which leads to reduced spectral reflectance. Significant variations in soil spectral reflectance are observed when the SMC is 16% or less. However, once the SMC exceeds 16%, the influence of moisture content on soil spectral reflectance diminishes, and the rate at which spectral reflectance decreases slows down. At SMCs of 20% and 22%, the corresponding curves are almost indistinguishable.

Figure 2. (a) The mean spectra for different SMCs. (b) The difference in average reflectance between dry soil samples and wet soil samples for different SMCs. (c) The average spectrum after continuum removal with different SMCs.

Principal Component Analysis (PCA) was conducted on the original spectra. The cumulative variance of dry soil and low SMC surpasses that of high SMC. Soil with a high
water content exhibited a notable decrease in PC1 and an increase in PC2, as indicated in Table 2.

Table 2. Explained and accumulation variance in the first two Principal Components.

<table>
<thead>
<tr>
<th>SMC</th>
<th>PC1 (%)</th>
<th>PC2 (%)</th>
<th>Accumulation (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dry soil</td>
<td>98.73</td>
<td>0.76</td>
<td>99.49</td>
</tr>
<tr>
<td>12%</td>
<td>97.90</td>
<td>1.35</td>
<td>99.25</td>
</tr>
<tr>
<td>14%</td>
<td>99.05</td>
<td>0.58</td>
<td>99.63</td>
</tr>
<tr>
<td>16%</td>
<td>98.27</td>
<td>1.03</td>
<td>99.29</td>
</tr>
<tr>
<td>18%</td>
<td>97.40</td>
<td>1.52</td>
<td>98.92</td>
</tr>
<tr>
<td>20%</td>
<td>95.34</td>
<td>3.11</td>
<td>98.45</td>
</tr>
<tr>
<td>22%</td>
<td>90.53</td>
<td>8.04</td>
<td>98.57</td>
</tr>
</tbody>
</table>

The diagram below clearly depicts differences in the spectra of dry soil, low SMC, and high SMC. The dry soil samples exhibit clear distinctions, whereas the scores for PC1 and PC2 of the other moist samples are mixed, making it challenging to discern clear clusters (Figure 3). Nonetheless, distinct clusters are still identifiable between low-moisture-content (12%, 14%, 16%) and high-moisture-content (18%, 20%, 22%) samples.

According to the calculation of the difference in average reflectance between the wet and dry soil samples (Figure 2b), the spectral absorption near 1950 nm is the highest among the full wavelength range and greater than other wavelengths. It can be observed from the graph that the influence of SMC on the spectral reflectance in the visible light range is significantly weaker compared to the near-infrared range. Figure 2c represents the continuum removal (CR) plot of the average spectral reflectance of dry soil samples and wet soil samples for seven different SMCs. The soil spectrum exhibits significant absorption around 1455 nm, 1945 nm, and 2200 nm.

Based on the analysis provided, there are notable differences in soil spectral reflectance across various soil moisture levels. Specifically, a pronounced water absorption valley near 1945 nm partially obscures the absorption valley near 2200 nm. This could potentially impact the prediction of soil organic, atter (SOM) content. Hence,
employing efficient SMC correction algorithms to calibrate spectra is necessary to mitigate moisture’s influence on SOM content prediction.

3.2. Construction of Spectral Index

Figure 4 illustrates the correlation coefficient curves between soil organic matter (SOM) content and different spectral transformations. SOM content demonstrates a significant negative correlation ($p < 0.01$) with the original spectrum in the 500–2400 nm wavelength range (Figure 4a). Conversely, it exhibits a significant positive correlation ($p < 0.01$) with the Log(1/R) spectrum, with correlation coefficients exceeding 0.3 in absolute value. Notably, the correlation in the 500–900 nm wavelength range surpasses that in other ranges, with correlation coefficients exceeding 0.55 in absolute value. Furthermore, the correlation coefficient of the Log(1/R) spectrum is even higher (Figure 4b). SOM content shows varying degrees of positive or negative correlations with the CR spectrum. When employing the original and Log(1/R) spectra to infer SOM content in the study area, the sensitive bands primarily reside in the visible and near-infrared bands. Conversely, if CR spectral data are utilized, the sensitive band range widens to cover visible light, near-infrared, and short-wave infrared band ranges (Figure 4c).

![Figure 4. Correlation coefficient curves between SOM content and reflectance under different spectral transformations: (a) R, (b) Log(1/R), and (c) CR.](image)

Within the full wavelength range, the sensitive bands for SOM content inversion are mainly located at 500–900 nm, 1450–1700 nm, and 2200–2400 nm. Based on our analysis of Figure 4, when studying the inversion of SOM content using the original spectrum and Log(1/R) spectrum, the sensitive bands are mainly in the visible light and near-infrared ranges. When using the CR spectrum for the inversion of SOM content in the study area, there are sensitive bands in the visible light, near-infrared, and short-wave infrared ranges, with a wider range of selection.

Figure 5 shows the contour maps of three types of correlation coefficients between SOM content and spectral indices in the study area, where different values represent the absolute correlation coefficients between $DSI$, $RSI$, $NDSI$, and SOM. The contour maps of the correlation coefficients clearly demonstrate the distribution range of the characteristic indices that have a relatively high correlation coefficient.

In the original spectral data, the band combinations exhibiting a high correlation between SOM content and $DSI$ are primarily situated in the range of 2160 to 2300 nm (Figure 5a). The range of band combinations showing a high correlation between SOM content and $RSI$ or $NDSI$ remains consistent and broader, spanning from 1400 to 2200 nm. Notably, $NDSI$ achieves the highest correlation coefficient of 0.9094, with a band combination of 2004 nm and 1900 nm (Figure 5b,c).
Figure 5. Correlation coefficient equipotential diagram between SOM content and three spectral indices.

In the Log(1/R) spectra (Figure 5d–f), the band combinations displaying a high correlation between SOM content and DSI are predominantly spread across ranges of 1400 to 1600 nm and 1900 to 2200 nm. The contour maps of correlation coefficients for RSI and NDSI exhibit similar distributions, with a narrower range of band combinations showing a high correlation primarily in the range of 2180 to 2280 nm. Notably, DSI achieves the highest correlation coefficient of 0.9081, with a band combination of 1900 nm and 2004 nm.

In the CR spectra (Figure 5g–i), the band combinations exhibiting a high correlation between SOM content and DSI are primarily concentrated in the narrow range of 2100 to 2300 nm. The contour maps of correlation coefficients for RSI and NDSI demonstrate similar distributions, with both being within the range of 2000 to 2280 nm. Notably, NDSI achieves the highest correlation coefficient of 0.9069, with a band combination of 2004 nm and 2180 nm.

Figure 4 shows that the correlation coefficients between reflectance at 2197 nm and 2266 nm and SOM range from −0.44 to −0.43 for the original spectra but significantly improve after calculating DI, RI, and NDI, with correlation coefficients exceeding 0.82 in absolute value. The Log(1/R) spectra exhibit a significantly improved correlation with SOM after calculating DI, RI, and NDI at 1900 nm, 2003 nm, and 2005 nm. Similarly, the CR spectra show a significantly improved correlation with SOM for spectral indices calculated at 2267 nm, 2202 nm, and 2200 nm. This indicates that after appropriate spectral index calculations, the spectral reflectance in different bands enhances the correlation with SOM, thereby improving the accuracy of modeling and prediction to a certain extent.
The sensitive bands of different spectral transformations and spectral indices and their correlations are summarized in Table 3.

### Table 3. Spectral index-sensitive bands and their correlations.

<table>
<thead>
<tr>
<th>Spectral Transformation</th>
<th>Spectral Index</th>
<th>Maximum Correlation Band Combination</th>
<th>Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>DSI</td>
<td>2267 2197</td>
<td>0.8851</td>
</tr>
<tr>
<td></td>
<td>RSI</td>
<td>2004 1900</td>
<td>0.9094</td>
</tr>
<tr>
<td></td>
<td>NDSI</td>
<td>2004 1900</td>
<td>0.9094</td>
</tr>
<tr>
<td>Log(1/R)</td>
<td>DSI</td>
<td>1900 2003</td>
<td>0.9080</td>
</tr>
<tr>
<td></td>
<td>RSI</td>
<td>2198 2267</td>
<td>0.8890</td>
</tr>
<tr>
<td></td>
<td>NDSI</td>
<td>2197 2266</td>
<td>0.8868</td>
</tr>
<tr>
<td>CR</td>
<td>DSI</td>
<td>2267 2201</td>
<td>0.8868</td>
</tr>
<tr>
<td></td>
<td>RSI</td>
<td>2005 1901</td>
<td>0.9076</td>
</tr>
<tr>
<td></td>
<td>NDSI</td>
<td>2005 1901</td>
<td>0.9069</td>
</tr>
</tbody>
</table>

The normalized difference spectral index based on reflectance \( R \) (NDSI\(_R\)) has a simple form and good performance, so a three-band normalized spectral index was constructed by adding a third band \( \lambda_k \) to the existing two-band spectral index, NDSI \( (\lambda_i, \lambda_j) \).

Ten two-band spectral indices with correlation coefficients above 0.7 with SOM were selected. A total of 10 sets of three-band spectral indices were constructed by iterating through the bands from 500 to 2400 nm and using Equation (6) to calculate every possible combination \((10 \times 1901)\). The three-band spectral index with the strongest correlation in each set was selected. As shown in Table 4, compared with the two-band spectral index, the vast majority of the three-band spectral indices exhibited an improved correlation with SOM.

\[
\text{NDSI}(\lambda_i, \lambda_j, \lambda_k) = \frac{R(\lambda_i) - R(\lambda_j) - R(\lambda_k)}{R(\lambda_i) + R(\lambda_j) + R(\lambda_k)}
\]  

(6)

### Table 4. Filtered three-band spectral indices.

<table>
<thead>
<tr>
<th>( \lambda_i ) (nm)</th>
<th>( \lambda_j ) (nm)</th>
<th>Correlation with Two Bands</th>
<th>( \lambda_k ) (nm)</th>
<th>Correlation with Three Bands</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>2380</td>
<td>0.91</td>
<td>1900</td>
<td>0.91</td>
</tr>
<tr>
<td>2300</td>
<td>2380</td>
<td>0.76</td>
<td>2380</td>
<td>0.77</td>
</tr>
<tr>
<td>2120</td>
<td>2360</td>
<td>0.72</td>
<td>2298</td>
<td>0.74</td>
</tr>
<tr>
<td>2260</td>
<td>2200</td>
<td>0.86</td>
<td>2270</td>
<td>0.89</td>
</tr>
<tr>
<td>2140</td>
<td>2200</td>
<td>0.74</td>
<td>2266</td>
<td>0.82</td>
</tr>
<tr>
<td>1360</td>
<td>1400</td>
<td>0.71</td>
<td>1509</td>
<td>0.88</td>
</tr>
<tr>
<td>1120</td>
<td>1100</td>
<td>0.71</td>
<td>1191</td>
<td>0.74</td>
</tr>
<tr>
<td>2140</td>
<td>1420</td>
<td>0.72</td>
<td>1485</td>
<td>0.80</td>
</tr>
<tr>
<td>1460</td>
<td>1440</td>
<td>0.85</td>
<td>1448</td>
<td>0.88</td>
</tr>
<tr>
<td>1540</td>
<td>1400</td>
<td>0.82</td>
<td>1342</td>
<td>0.85</td>
</tr>
</tbody>
</table>

When selecting spectral indices, it is crucial to ensure that the input matrix for EPO is not too small. Hence, from the nine different spectral transformations and operations with a correlation with SOM above 0.82 (Figure 6), collinearity was evaluated using the Variance Inflation Factor (VIF). Subsequently, a total of 190 spectral indices, comprising 180 two-band and 10 three-band indices, were selected to replace the wet soil reflectance matrix as the input matrix for both the EPO and DS methods.
Figure 6. Spectral index equipotential diagram with SOM correlation degree above 0.82.

3.3. EPO-PLS and DS-PLS Prediction Results after Using Spectral Index

3.3.1. PLS Prediction of Wet Soil and Dry Soil

When no soil moisture correction was applied, the $R^2$ for wet soil was 0.5259, much lower than the $R^2$ for dry soil (0.76), and $RMSEP$ was 4.6433 (Figure 7a,b). The high prediction error for wet soil using PLS prediction indicates that the prediction accuracy is significantly reduced when soil samples contain a certain amount of water, and the model cannot accurately estimate SOM under varying soil moisture conditions. This indicates that the model is unable to accurately estimate SOM under varying soil moisture conditions, underscoring the necessity to develop an effective soil moisture correction algorithm.
3.3.2. Model of Spectral Index Combined with EPO

In this model, the dimensions \( i \) of EPO and the optimal number of factors \( j \) in PLS need to be further optimized to improve the prediction accuracy of the EPO-PLS model. The \( RMSE \) value and \( R^2 \) in the model are used as criteria to determine the best parameter combination. The cross-validation \( RMSE \) curves generated by different combinations of \( i \) and \( j \) in the EPO-PLS model are shown in Figure 8a. The optimization range for dimension \( i \) in the EPO-PLS model is 1 to 13, and the optimal range for the number of factors \( j \) is 1 to 13. The results show that there is no significant decrease in \( RMSE \) value after four PLS factors, and the lowest \( RMSE \) value is achieved when the EPO dimension is 4. The correlation coefficient error (\( R^2 \)) curve is shown in Figure 8b, and the model’s \( R^2 \) value reaches the maximum when \( i = 4 \) and \( j = 4 \). Therefore, \( i = 4 \) and \( j = 4 \) is the best parameter combination for subsequent modeling analysis.

After SMC correction by EPO, the \( R^2 \) of the EPO-PLS model is 0.63, and the \( RMSE_{p} \) is 4.1021, indicating a moderate accuracy for the model. This model can be used to roughly estimate SOM under different moisture gradients (Figure 9a). Compared with Figure 7b, the prediction accuracy has been improved after correction, with an increase in the \( R^2 \) of 0.1041,
indicating the feasibility of using EPO to remove the influence of SMC and improve the accuracy of SOM estimation. The $R^2$ of the SI EPO-PLS model, which combines spectral indices and EPO for modeling analysis, is 0.735, and the $\text{RMSEP}$ is 3.4102. This model demonstrates the capability to accurately estimate SOM under varying moisture gradients, showcasing good model accuracy, low bias, and strong explanatory power (Figure 9b).

**Figure 9.** SOM prediction scatter plot: (a) EPO-PLS; (b) SI EPO-PLS.

### 3.3.3. Model of Spectral Index Combined with DS

To further verify the superiority of spectral index processing, an analysis was conducted on the combination modeling of spectral indices with the DS, as shown in Figure 10. Compared to the DS-PLS model, the SI DS-PLS model, which incorporates spectral indices, showed an $R^2$ improvement of 0.09 and a decrease in $\text{RMSEP}$ of 0.5721, indicating an enhancement in prediction accuracy. However, there is still a gap compared to the prediction results for dry soil ($R^2 = 0.76$, $\text{RMSEP} = 3.0161$). When comparing the prediction accuracy of EPO-PLS (Figure 9a) and DS-PLS (Figure 10a), it can be inferred that the effectiveness of EPO in water removal surpasses DS in small-sample prediction.

**Figure 10.** SOM prediction scatter plot: (a) DS-PLS; (b) SI DS-PLS.
3.4. EPO-PLS Prediction Results after Using Three-Band Spectral Index

Comparing the effects of two different removal methods, full-band EPO and DS, on the estimation of SOM, it can be concluded that both SMC correction methods can alleviate the influence of soil moisture. However, in small-sample prediction, the performance of EPO ($R^2 = 0.63$, $RMSEP = 4.1021$) is superior to DS ($R^2 = 0.61$, $RMSEP = 4.2806$). Therefore, it is worth discussing whether the accuracy of SOM prediction can be further improved by discussing EPO combined with three-band spectral indices.

As mentioned above, the spectral index of the two bands uses three forms, of which the normalized form is the best. We apply the best of this form to three bands. This normalized three-band spectral index is not unique, so we introduce six coefficients.

The constructed spectral index form is as follows:

$$\text{NDSI}(\lambda_1, \lambda_2, \lambda_3) = \frac{aR(\lambda_1) + bR(\lambda_2) + cR(\lambda_3)}{dR(\lambda_1) + eR(\lambda_2) + fR(\lambda_3)}$$

(7)

The full spectrum (500–2400 nm) was traversed with band intervals of 5 nm and 10 nm. Spectral indices whose correlation with SOM was greater than 0.75 were screened and ranked by number. The top nine of them were selected and used to construct the three-band spectral indices in this paper. The resulting parameters are shown in Table 5.

Table 5. The filtered three-band spectral indices parameters.

<table>
<thead>
<tr>
<th>Number</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDSI 1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>NDSI 2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>NDSI 3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>NDSI 4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>NDSI 5</td>
<td>1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>NDSI 6</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>NDSI 7</td>
<td>1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>NDSI 8</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>NDSI 9</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

It is worth noting that the nine initial spectral indices, arranged by their correlation coefficients with organic matter exceeding 0.75, remain consistent for both 5 nm and 10 nm intervals. This suggests that the bands sensitive to SOM constitute a distinct range, and the selection of wavelength intervals within this range is not significantly sensitive.

Figure 11 shows the contour map of the correlation coefficients between SOM content and nine three-band spectral indices in the study area. The contour map of the correlation coefficients clearly demonstrates the distribution range of characteristic indices with higher correlation coefficients.

In the selection of spectral indices, the input matrix for EPO should not be too small. Therefore, among the 9 three-band NDSIs with a correlation coefficient with SOM above 0.75, collinearity was detected using the Variance Inflation Factor (VIF). A total of 180 spectral indices were selected to replace the reflectance matrix as the input matrix for EPO. The predicted map obtained after removing the moisture information using EPO is shown in Figure 12.
Figure 11. The contour map of the correlation coefficients between SOM content and 9 three-band spectral indices.

Figure 12. SOM prediction scatter plot: SI3 EPO-PLS.

The $R^2$ value for the combined modeling analysis (SI3 EPO-PLS) utilizing three-band spectral indices and EPO is 0.752, with an RMSEp of 3.1344. Compared to SI EPO-PLS, the $R^2$ increases by 0.022, and the RMSEp decreases by 0.3788. In comparison to EPO-PLS prediction, the $R^2$ increases by 0.122, and the RMSEp decreases by 0.9677. This suggests that the combination of three-band spectral indices and EPO enables a more precise estimation of SOM across varying SMCs. This highlights the superior capacity of three-
Remote Sens. 2024, 16, 2065

4. Discussion

4.1. Effect of SMC on Spectrum

When soil moisture content (SMC) is equal to or less than 16%, variations in soil spectral reflectance are notably influenced by changes in moisture content. However, when SMC exceeds 16%, the decrease in soil spectral reflectance occurs at a slower rate. Notably, between 20% and 22% SMC, the two curves nearly coincide (Figure 2). This observation indicates that at higher SMC levels, spectral reflectance becomes less sensitive to changes in moisture content. Principal Component Analysis (PCA) was conducted on the original spectra as well. Principal Component 1 and Principal Component 2 (PC1, PC2) collectively account for over 90% of the spectral variance within the dataset (Figure 3). Notably, dry soil samples exhibit clear differentiation, while the scores for the moist samples show overlapping PC1 and PC2 values, failing to form distinct clusters. However, clusters corresponding to low SMC levels (12%, 14%, 16%) and high SMC levels (18%, 20%, 22%) are discernible. Moisture alters the physical structure and color of the soil, resulting in diminished soil spectral reflectance. At a higher moisture content, changes in color become less pronounced. As soil moisture increases, water absorption by soil particles fills both micropores and macropores, altering the relative refractive index of the soil particle surface. However, this has minimal impact on reflectance when water covers these pores. This observation aligns with the findings of Yudina [34].

Therefore, an efficient and useful soil moisture elimination algorithm must be used to correct the spectra and reduce the impact of SMC on SOM prediction.

4.2. The Influence of the Spectral Index on the Correlation between Spectrum and SOM

Detecting soil organic matter content is pivotal for assessing soil fertility. Presently, most research relies on full-spectrum spectral data to predict soil properties. However, soil hyperspectral curves encompass multiple bands, leading to varying degrees of correlation among them and resulting in information redundancy [35]. In the spectral detection of SOM, extracting sensitive bands through spectral preprocessing is crucial to reduce redundant information from full-band modeling. This study focuses on extracting such sensitive bands of SOM. By utilizing soil spectral indices, the input variables for soil attribute spectral modeling can be streamlined, enhancing modeling efficiency. In summary, the one-band spectral index suits simple applications, the two-band spectral index offers more informative features, and the three-band spectral index provides comprehensive analysis capabilities. Moreover, this study visualizes the correlation between common spectral indices and measured properties in a visual format, taking into account the interrelationships between spectra.

This study found that the sensitive bands for the inversion of SOM content are mainly located at 500–900 nm, 1450–1700 nm, and 2200–2400 nm (Figure 4), which corresponds to previous research [36]. After calculating spectral reflectance in various bands using suitable characteristic indices, the correlation between spectral data and SOM was found to be strengthened, thereby enhancing modeling prediction accuracy to some degree. This paper introduces a method for constructing a three-band spectral index which shows a considerable improvement in correlation with SOM compared to the two-band spectral index. The directly constructed three-band spectral index exhibits a stronger correlation with SOM (Figure 11). The information of the three bands provides more comprehensive data characteristics, but it needs to process more band information, which requires a large amount of calculation and requires more careful screening.
4.3. The Model Advantage of Spectral Index Combined with Water Removal Algorithm

This study combined EPO with spectral index combination modeling, which can effectively deal with nonlinear influencing factors in different SMCs and improve the accuracy of SOM estimation (Figure 9). Compared with the EPO-PLS model, the $R^2$ of the SI EPO-PLS model increased by 0.1, and the RMSE decreased by 0.7919. $R^2$ is 0.745, which is close to the prediction result for dry soil in Figure 7b ($R^2 = 0.76$, RMSE = 3.0161). The model has been improved from the usable level to the better level, which shows that the method of combining EPO and the spectral index can effectively handle the nonlinear influencing factors in different SMCs and improve the accuracy of SOM. In order to further verify the excellence of spectral index processing data, the method of combining DS and the spectral index was analyzed. This shows that the spectral index, combined with other algorithms, can be used to deal with non-linear influencing factors in different SMCs and further improve SOM accuracy, and it may be generalizable (Figure 10). Comparing the prediction accuracy of EPO-PLS (Figure 9a) and DS-PLS (Figure 10a) shows that the ability of EPO to remove SMC is better than that of DS in small-sample prediction.

At the same time, this method is comparable to other SOM prediction methods. In 2023, Cao proposed a dynamic normalized difference index method, which improved the $R^2$ of organic matter prediction by 0.14 [37]. In the same year, Wu proposed the SESMRT model, which improved the $R^2$ by 0.0064 [38]. The Kubelka–Munk theory-based spectral correction model for soil moisture removal proposed by Ou increased the $R^2$ by 0.22 [39]. In 2024, Wu proposed an RF machine learning model combined with CARS selected feature bands to increase the $R^2$ by 0.2 [40]. Comparing with other techniques published in recent years, the accuracy of SOM prediction by this method is higher than the average level ($R^2$ increased by 0.23).

To further enhance the pretreatment method for spectral reflectance and improve the accuracy of SOM estimation, future research could focus on establishing an effective physical model combined with statistical methods or other physical models [41]. Additionally, expanding the sample set used in this study could provide more comprehensive insights for further in-depth discussion.

5. Conclusions

This study explores the feasibility of using two-band and three-band mixed spectral indices, a three-band spectral index, and two moisture removal algorithms (EPO and DS) to estimate SOM in soil affected by moisture. As shown in previous research results, the spectral reflectance decreases nonlinearly with the increase in water content, and the SOM estimation performance decreases significantly ($R^2: 0.76–0.5259$, RMSEP: 3.0161–4.6433 g/kg). The spectral index after spectral transformation can effectively improve the correlation between spectral information and SOM. In general, the estimation accuracy of the spectral index combined with the water removal algorithm ($R^2 = 0.752$, RMSEP = 3.1344 g/kg) is better than the estimation accuracy of the full-spectrum data ($R^2 = 0.63$, RMSEP = 4.1021 g/kg). The three-band spectral index is better than the two-band and three-band mixed spectral indices ($R^2 = 0.735$, RMSEP = 3.4102 g/kg). In addition, we also compared the effects of two different removal methods (EPO and DS), based on full-band modeling, on soil SOM estimation. It can be concluded that both of the soil moisture correction methods tested can alleviate the impact of soil moisture. However, in small-sample prediction, the performance of EPO is higher than that of DS ($R^2 = 0.61$, RMSEP = 4.2806 g/kg).

Author Contributions: Conceptualization, J.X.; methodology, J.X.; software, J.X.; validation, J.X., J.Y. and C.Y.; formal analysis, J.X.; investigation, J.X.; resources, J.X. and Y.L.; data curation, J.X.; writing—original draft preparation, J.X.; writing—review and editing, C.Y., J.Y. and Y.L.; visualization, J.X. and J.Y.; supervision, C.Y. and J.Y.; project administration, C.Y.; funding acquisition, C.Y. All authors have read and agreed to the published version of the manuscript.
Funding: This research was funded by the Jilin Key R&D Program of China under Grant 20230201036GX; it was also funded in part by the National Natural Science Foundation of China (NSFC) under Grant 62105331 and Grant 62275114.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Acknowledgments: The editor and the reviewers are thanked for their helpful comments and criticisms of the initial draft of this paper.

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

References


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.