Article

Using Optimized Spectral Indices and Machine Learning Algorithms to Assess Soil Copper Concentration in Mining Areas

Chang Meng [1,2], Mei Hong [1,2,*], Yuncai Hu [3] and Fei Li [1,2,*

[1] Inner Mongolia Key Laboratory of Soil Quality and Nutrient Resource, Hohhot 010018, China; 2021202040012@emails.imau.edu.cn
[2] Key Laboratory of Agricultural Ecological Security and Green Development, Universities of Inner Mongolia Autonomous, Hohhot 010018, China
[3] Precision Agriculture Laboratory, School of Life Sciences, Technical University of Munich, 85354 Freising, Germany; yc.hu@tum.de
* Correspondence: nmczhm1970@126.com (M.H.); lifei@imau.edu.cn (F.L.)

Abstract: Soil copper (Cu) contamination in mining areas poses a serious threat to the surrounding environment and human health. Timely determination of Cu concentrations is crucial for the ecological protection of mining areas. Hyperspectral remote sensing technology, with its non-destructive monitoring advantages, is essential for monitoring soil Cu pollution and achieving sustainable agricultural development. Using the hyperspectral technique for assessing soil Cu concentration, four machine learning models (support vector regression (SVR), random forest (RF), partial least squares regression (PLSR), and artificial neural network (ANN)), combined with three types of input variables (the full-band, sensitive bands, and optimized spectral indices (Opt-TBIs)) were employed. The hyperspectral reflectance of 647 soil samples from an abandoned tailings mine in western Inner Mongolia, China was collected. The sensitive bands were extracted using the successive projections algorithms (SPA), and 12 Opt-TBIs were selected. Results showed that the regions with higher soil Cu concentration extracted by SPA and Opt-TBIs were concentrated in the red edge and near-infrared regions. Compared with the full spectrum and SPA-sensitive bands, models based on Opt-TBIs successfully predicted soil Cu concentrations. The Opt-TBIs-RF model provided higher accuracy in estimating soil Cu among the four models. Using only four Opt-TBIs as input variables, the model maintained a stable performance in estimating Cu concentrations in different mining areas (R^2_{Val} = 0.72, RPD_{Val} = 1.90). In conclusion, Opt-TBIs as input variables demonstrate good predictive capabilities for soil Cu concentrations in the study area, providing a basis for the formulation of sustainable strategies for soil reclamation and environmental protection in Inner Mongolia.

Keywords: soil Cu concentration; machine learning; optimized two- and three-band spectral indices (Opt-TBIs); hyperspectral monitoring technique; soil environmental protection

check for updates

Citation: Meng, C.; Hong, M.; Hu, Y.; Li, F. Using Optimized Spectral Indices and Machine Learning Algorithms to Assess Soil Copper Concentration in Mining Areas. *Sustainability* **2024**, *16*, 4153. https://doi.org/10.3390/su16104153

Academic Editors: Anna De Marco and Jiulan Dai

Received: 26 February 2024
Revised: 29 April 2024
Accepted: 13 May 2024
Published: 15 May 2024

1. Introduction

The large-scale mining of copper (Cu) ore brings economic benefits. However, the resulting massive amounts of slag and waste rock are not effectively utilized or managed, ultimately accumulating into tailings ponds. Prolonged accumulation dramatically increases Cu concentration in the surrounding soil, far exceeding its natural carrying capacity [1–3]. Excessive Cu in the soil is absorbed and accumulated by passive uptake of plants, ultimately leading to human Cu poisoning and causing liver and kidney failure. Specifically, high Cu concentrations induce toxicity in plants, e.g., disrupting the root structure and reducing plant growth [4]. Cu can also alter the quantity and community of soil microorganisms, causing changes in the soil's physicochemical properties [5]. Ultimately, it alters soil structure and function, posing significant threats to the surrounding ecological environment [6]. Therefore, there is an urgent need for reliable monitoring tools to rapidly and accurately assess the soil Cu in mining areas, providing adequate scientific evidence for regional heavy metal Cu pollution control and sustainable soil management.

With hyperspectral remote sensing technology's continuous advancement and development, its application to soil organic matter, heavy metals, and other aspects has become increasingly widespread [7]. Hyperspectral data accurately reflect soil spectral characteristics, which is crucial in extracting spectral information and establishing models. Hyperspectral remote sensing technology, with its high resolution, wide spectral range, and fast analysis speed, demonstrates significant advantages in monitoring soil heavy metal concentrations [8–10]. Compared to traditional field sampling and laboratory wet chemical measurements, this technology can rapidly and efficiently obtain soil Cu concentrations over large areas [11]. However, there are numerous hyperspectral bands in soil, and the spectral information is complex, with a considerable amount of redundant information between spectral variables. If all spectral bands are used for modeling, the inversion model calculation is affected by redundant information interference, reducing the accuracy and timeliness of model predictions. Therefore, extracting two or more strong information wavelength combinations of spectral indices is necessary to effectively reduce irrelevant information interference, enhancing model computational efficiency and accuracy [12]. Previous studies have shown that spectral index formulas such as ($R_1 - R_2$) and ($R_1 - R_2$)/sqrt($R_1 + R_2$) have been well utilized for estimating metal concentrations in plants and minimizing spectral shifts caused by external factors [13]. However, the sensitivity to high-concentration heavy metals using only two-band spectral indices is relatively low because two-band spectral indices become saturated with increasing metal concentrations. We introduce three-band spectral indices to address this, which can somewhat reduce spectral saturation limits [14]. Shi [15] pointed out that the three-band indices are a more accurate estimation than the two-band indices because the three-band indices involve more reasonable and informative bands. Many studies have also proved that the three-band spectral indies have apparent advantages in estimating heavy metal concentration [16–18]. In recent years, the studies on the migration and transformation laws between plants and soil heavy metals have gradually deepened, with numerous spectral index formulas gradually being introduced into soil heavy metal band optimization. For example, Kooistra [19] found that (R_1-R_2) can reasonably estimate soil heavy metal concentration, and R^2 ranges from 0.50 to 0.73. Jiang's [20] work also confirmed that using a combined spectral index method significantly enhances the correlation between spectral variables and soil Cu concentrations. To further improve the reliability of spectral index extraction bands, using fractional derivatives combined with band combination algorithms effectively mines more soil Cu spectral information [21]. In addition, the three-band metal element index (TSMEI) developed by Fu [22] can better monitor arsenic concentration in soil. However, due to the complexity of soil composition, low Cu concentrations, and weak spectral information, it is still unclear which spectral indices are the most effective for estimating soil Cu. Therefore, combining TBIs with machine learning models is necessary to improve the accuracy of model monitoring of heavy metals.

Applying machine learning algorithms combined with hyperspectral remote sensing data in soil heavy metal detection has become increasingly popular. Tan [23] used the coupling of 2151 hyperspectral bands with an RF algorithm to predict the soil heavy metal of Cu concentration with a lower coefficient of determination. Similarly, the research findings of Cheng [24] confirmed that the PLSR model could predict 50–70% of the heavy metal variations in soil using full-band analysis. However, there are many hyperspectral data channels, most of which have nothing to do with the required elements. Accordingly, most of the uncorrelated spectral information affects the estimation performance of hyperspectral models [25,26]. Therefore, choosing the appropriate band as the input variable is significant for the machine learning algorithm. There are many methods for variable selection, and one method commonly used for dimensionality reduction is the successive projection algorithm (SPA). The SPA selects a group of valuable features containing helpful information and the most important data of the original dataset. Wang et al. [27] used the SPA combined with PLSR to invert 89% of soil salinity changes in saline–alkali land. Similarly, Peng [28] found that the SPA combined with support vector regression could explain 61% of soil organic matter concentration variation. Currently, SPA methods are primarily used in plant and chemical studies. There are few applications in soil heavy metal research. Another critical method is determining how to select spectral indices. Many studies have shown that constructing models based on spectral indices can complement information between bands. The model's predictive accuracy and stability can be enhanced by selecting spectral bands with the highest correlation to soil Cu concentrations [29,30]. For example, the optimized combination of $(R_1 - R_2)/(R_1 + R_2)$ with PLSR can explain 85% of cadmium concentration variations [31]. Combining the RF algorithm with spectral indices significantly enhances the accuracy of predicting soil heavy metal Cu concentrations [32]. Using TBIs to select the optimal bands for predicting soil Cu concentrations compensates for the limitations of using full bands to predict soil heavy metal Cu concentrations. However, although these band selection methods can be used for feature extraction, accurate model estimation can only be achieved by combining the best bands as input variables. Therefore, the performance of soil heavy metal estimation using different input variables and machine learning combinations still warrants further investigation.

Machine learning methods show the advances in monitoring heavy metals in soil, but the process of estimating soil heavy metal concentrations is strongly influenced by mathematical models. Due to variations in model function formulas and data processing methods, there are differences in input variables when combining different machine learning approaches. Therefore, the main objective of this study was (1) to evaluate the performance of spectral indicators in estimating soil Cu concentration and compare the effects of different input variables (full-band, SPA, spectral indices) on the accuracy of soil Cu prediction, and (2) to analyze the prediction accuracy and stability of the model and select the appropriate model and input variables.

2. Materials and Methods

2.1. Study Area

Figure 1 shows the location of the soil sample collection in the study area. Figure 1a is located near an abandoned Cu tailings pond in western Inner Mongolia, China. The terrain is high in the southwest and low in the northeast, slightly inclined from the southwest to the northeast, and located in the Loop Plain. The average annual temperature is 3.7–7.6 °C, yearly sunshine hours are 3213.7 h, annual rainfall is 136.3 mm, and maximum permafrost depth is 1.27 m. The Cu contaminated soil in this area is predominantly saline–alkali soil, with abundant mineral and sunlight resources within its boundaries.

The ore body in the Cu tailings pond area mainly exhibits a layered structure, with the ore structure predominantly in the form of veins and blocks. The mineral content primarily consists of hematite and magnetite, with small amounts of calcium carbonate and trace amounts of aluminosilicate. Iron and manganese oxides in the soil, and carbonates and silicates, can form Cu compounds through adsorption processes.

To validate the stability of the soil Cu concentration model constructed for the abandoned Cu tailings pond, the region depicted in Figure 1b was selected as the validation area. Figure 1b is in a northern region of Inner Mongolia, near a tailings pond area. Climatic conditions vary with elevation, with temperatures decreasing from southeast to northwest. The average annual temperature ranges from 1–5 °C, with a growing season of 90–120 days. Precipitation is relatively low and unevenly distributed temporally and spatially, with considerable interannual variability. The average annual precipitation is around 370 mm.

The ore structure in the tailings pond area depicted in Figure 1b primarily exhibits a layered or pseudo-layered formation with an east–west orientation. The mineral content is predominantly composed of pyrite and silicate minerals, but a smaller proportion of magnetite. Cu compounds mainly consist of oxides and silicate compounds.

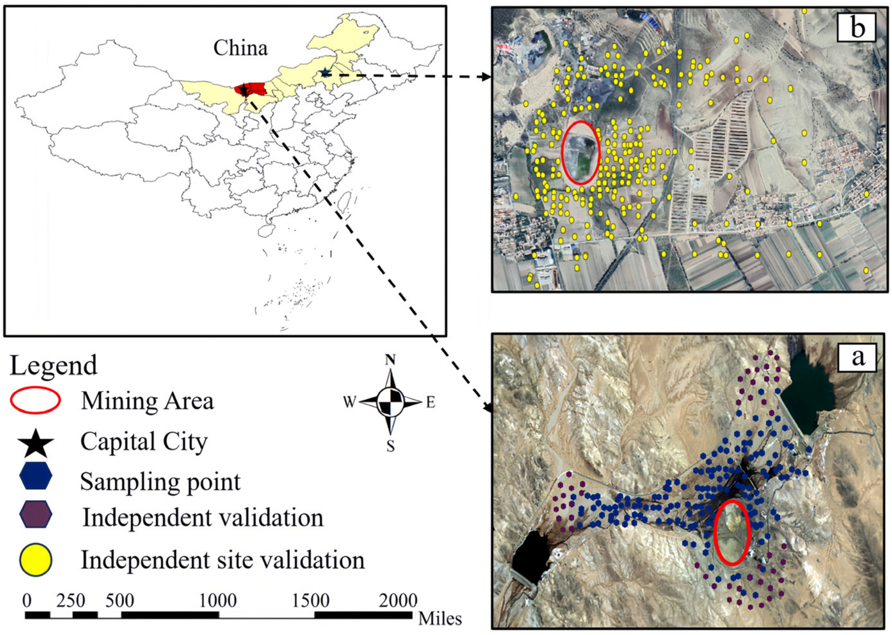


Figure 1. Distribution of sampling sites.

2.2. Data Collection

2.2.1. Soil Sample Collection and Measuring

Sampling was conducted in July 2020 around the western and northern tailings pond areas according to topography, wind direction, and water flow direction. The plum method was employed, and soil samples were collected at actual distances of 20, 50, 100, 150, 200, and 300 m, with 647 and 232 soil samples in the western and northern tailings pond areas, respectively. Nine sub-samples were combined to form one sample point. The collected soil samples underwent natural air-drying treatment under sunlight to avoid interference from soil moisture and other factors [33]. After drying, the soil was sieved to remove plant roots and large particles of sand and gravel. A 2 mm nylon sieve was then selected for sieving all soil samples. Subsequently, the soil was ground using an agate mortar and pestle until passing through a 0.15 mm sieve. The sieved soil samples were uniformly divided into two parts using the quartering method, with one part used for laboratory chemical analysis and the other part used for spectral collection.

The determination of soil samples primarily utilized the HF-HNO_3-HCl-$HClO_4$ four-acid stepwise microwave digestion method. Three parallel samples were set for each soil. Each soil sample weighed 0.15 g and was moistened with 0.5 mL of ultra-pure water, followed by adding 2 mL of hydrofluoric acid (HF), and left to stand for 12 h. Then, 6 mL of hydrochloric acid (HCl) and 2 mL of nitric acid (HNO_3) were added, and the soil digestion was carried out using a microwave digestion instrument. After digestion, 1 mL of perchloric acid ($HClO_4$) was added, and the mixture was placed in a fume hood for acid evaporation. Acid evaporation was carried out for at least 3 h. When only one drop of liquid remained in the digestion vessel, a small amount of 2% nitric acid was added while still warm to rinse the walls of the vessel. After cooling, the solution was made up to 50 mL in a volumetric flask, filtered through a 0.22 µm filter, and then was ready for analysis. The concentration of Cu was carried out using inductively coupled plasma–mass spectrometry (ICP-MS), with the accuracy and precision of the sample analysis method controlled by the National First-Level Soil Standard Substance (GSS-18).

The ICP-MS instrument model ICAP RQ (Thermo Fisher Scientific Inc., Waltham, MA, USA) was used for soil sample analysis. Before measuring the samples, standard curves of soil Cu were prepared at concentrations of 0, 20, 50, 100, and 300 µg/L. After the instrument was started for 30 min, the standard series was then sequentially introduced into the nebulizer, from low to high concentrations, for analysis, with the Cu mass concentration as the abscissa and the ratio of corresponding response values to internal standard values as the ordinate to establish the standard curve. Before measuring each sample, the system was rinsed with a 2% nitric acid solution until the signal decreased to the lowest level, and sample measurement began after the analysis signal stabilized. Laboratory blanks and national standards were prioritized for measurement before each sample was measured sequentially. The Cu concentration (ω_1, mg/kg) was calculated according to Formula (1).

$$\omega_1 = \frac{(\rho - \rho_0)\times V \times f}{m\times W_{dm}} 10^{-3} \tag{1}$$

where ω_1 is the Cu concentration in soil samples; ρ is the corresponding Cu concentration in the sample calculated from the standard curve; ρ_0 is the corresponding Cu element concentration in the blank sample; v is the final volume of the digested sample; f is the dilution factor of the sample; m is the mass of the sieved soil sample taken; W_{dm} is the content of sample dry matter.

2.2.2. Spectral Acquisition and Processing

In this study, soil spectral measurements were conducted in a completely dark laboratory to avoid interference from external light sources. Each soil sample was taken from an area of 20 cm^2 and placed in a specialized square container (with dimensions of 20 cm^2, non-reflective). A PSR+ ultra-portable full-band geophysical spectrometer (Spectral Evolution) covering a 350–2500 nm spectral wavelength was used. The measurement was performed using the instrument's own 100 W halogen lamp as the only light source. A fiber-optic probe was used to keep the probe perpendicular to the sample, with the lower end positioned 20 cm away from the sample. The angle and distance between the light source and the sample were set at 45° and 35 cm, respectively. A standardized white panel was used for calibration before the first scan to provide accurate measurements, and each sample was averaged after 10 repeated measurements. We randomly divided 647 soil spectral data into calibration and validation datasets. The calibration points comprised 90% (70% calibration and 20% validation) of the entire dataset, while the independent validation points were represented by the remaining 10% (Figure 1). The 232 soil data points from the tailings pond area depicted in Figure 1b were utilized to validate the model's performance in estimating soil Cu concentrations in different mining areas.

2.3. Methods

2.3.1. Successive Projections Algorithm (SPA)

The SPA can find the band containing the most information from the spectral information to reduce the collinearity between band information. The SPA can extract the maximum information bands from spectral data, reducing band collinearity. By projecting wavelengths onto others and comparing the magnitude of projection vectors, the wavelength with the maximum RMSE value is selected as the final feature wavelength. The SPA generally selects wavelength variable combinations with the most minor redundant information or minimum collinearity, effectively improving model computational efficiency and accuracy. Therefore, the SPA can reduce the original spectral information and solve the collinearity problem well.

In this study, feature band extraction was performed using the SPA (Figure 2), and the SPA was used to screen feature wavelengths for the original spectrum (Figure 2a). With the increase in the number of screening variables, the RMSE first decreased rapidly, and when the number of variables was 8, the RMSE tended to a stable state. Its value was 162.45 mg/kg (Figure 2b). Eight characteristic wavelengths were obtained by SPA operation, from only 0.03% of the whole band, which drastically reduced many redundant information variables in the spectral information.

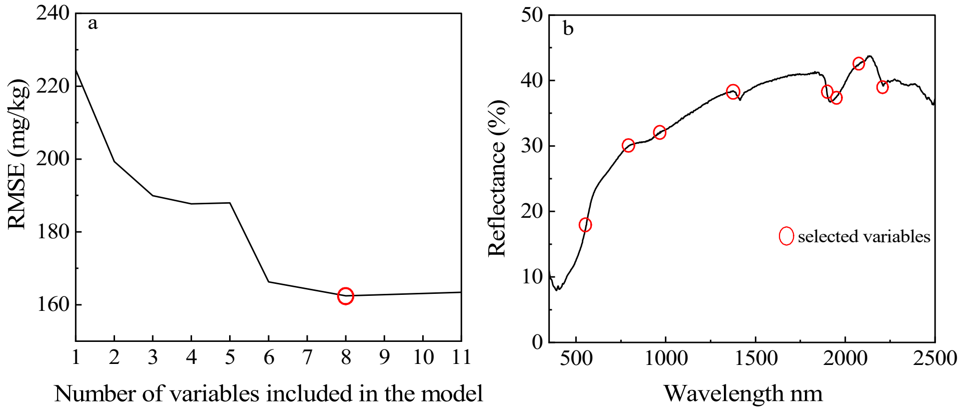


Figure 2. Successive projection algorithm: (**a**) final number of the eight selected variables (RMSE = 162.45 mg/kg), and (**b**) eight feature series.

2.3.2. Hyperspectral Indices

The application of spectral indices can reduce the sensitivity of irrelevant spectral bands to soil Cu elements, thereby mitigating redundancy in hyperspectral bands [34–39]. To investigate the impact of different types of spectral index formulas on the estimation of soil Cu concentration, this study selected 12 representative two-band and three-band spectral index (TBI) formulas (Table 1) and studied the wavelengths (R_1, R_2, and R_3) of TBIs within the range of 350–2500 nm. Utilizing TBI formulas, the optimal soil Cu bands were selected to construct optimized spectral indices (Opt-TBIs) for identifying soil Cu-sensitive bands. Opt-TBIs can improve the accuracy of soil Cu estimation to some extent by identifying the optimal combinations of sensitive bands. The process of extracting characteristic wavelengths of soil Cu involved combining Cu concentrations with each index formula, selecting the set of all band combinations with the most minor error between the calibration set and the validation set, and constructing spectral indices for the new band combination. The R^2 value consistently increased for each index formula, always retaining the top 1% of band combinations. The contour and slice correlation maps visualize the optimal wave combinations and sensitive spectral response regions. We used MATLAB 2021b to develop 2D and 3D correlation potentials of Opt-TBIs and Cu concentrations in soil.

Table 1. List of the hyperspectral indices used in this study.

Two-Band Optimized Spectral Indices		Three-Band Spectral Indices			
Spectral Indices	**Formulas**	**Spectral Indices**	**Formulas**	**Spectral Indices**	**Formulas**
TBI_1 (R_1, R_2)	$(R_1 - R_2)/(R_1 + R_2)$.	TBI_5 (R_1, R_2, R_3)	$(R_1 - R_2)/(R_1 - R_3)$	TBI_9 (R_1, R_2, R_3)	$(R_1 - R_2)/ (R_1 + R_2 - 2 \times R_3)$
TBI_2 (R_1, R_2)	$R_1 - R_2$	TBI_6 (R_1, R_2, R_3)	$R_1/(R_2 \times R_3)$	TBI_{10} (R_1, R_2, R_3)	$(R_1 - R_2)/(R_2 - R_3)$
TBI_3 (R_1, R_2)	$(R_1 - R_2)/\text{sqrt} (R_1 + R_2)$	TBI_7 (R_1, R_2, R_3)	$(R_1 - R_2)/R_3$	TBI_{11} (R_1, R_2, R_3)	$(R_1/(R_2 + R_3)$
TBI_4 (R_1, R_2)	$(R_1 + R_2)/R_2$	TBI_8 (R_1, R_2, R_3)	$(R_1 - R_2)/(R_3 - R_2)$	TBI_{12} (R_1, R_2, R_3)	$(R_1 + R_2)/R_3$

2.3.3. Machine Learning Algorithms

Partial Least Squares Regression (PLSR)

Machine learning algorithms are fundamental in hyperspectral analysis of soil heavy metal concentrations, especially in improving the performance of specific algorithms through empirical learning. The partial least squares regression (PLSR) model operates by detecting linear combinations of explanatory variables (controlled variables) to run processes or systems. The purpose of PLSR is to minimize the number of residual matrices in the response variable, while maintaining the correlation between the explanatory variables and the response variable through internal relationships unaffected by harmful collinearity in the explanatory variables [40]. In this study, the independent variable matrix is X and the dependent variable matrix is Y. The matrices of the standardized independent and dependent variables are represented as E and F, respectively. The regression variances of the first principal components t1 and u1 of the independent and dependent variables are solved, and the residual matrices E1 and F1 are computed. E1 and F1 are then used to replace E and F, forming new independent and dependent variables. The first principal components t2 and u2 of the new independent and dependent variables are solved and set as the second principal components of the original independent and dependent variables. New independent variables E1 and dependent variables F2, along with the regression equation of the second principal components t2 and u2, are established. These steps are repeated until all principal components are obtained. Cross-validation is performed to determine the number of principal components that meet the conditions, and a regression model is established. The procedures above were carried out for regression calculations of the model in a Python 3.10 environment, utilizing the "PLSR egression" function from the "sklearn" package for relevant operations of the PLSR model.

Support Vector Regression (SVR)

SVR is widely used as a model in classification and regression analysis, and is defined as a linear classifier with the most considerable interval in the feature space. The advantage of SVR is its good intrinsic generalization ability to handle high-dimensional input spaces. As a small-sample learning method, it simplifies the usual classification and regression problems. The SVR model selected in this study uses the kernel function from the "support vector regression" function in the "sklearn" package, implemented for model calibration and validation in a Python 3.10 environment. The specific process is shown in Figure 3.

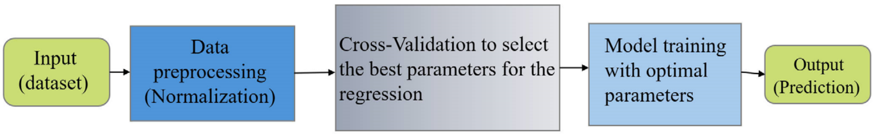

Input (dataset)
Data preprocessing (Normalization)
Cross-Validation to select the best parameters for the regression
Model training with optimal parameters
Output (Prediction)

Figure 3. SVR model flow chart.

Artificial Neural Networks (ANNs)

An ANN is a parallel information processing method consisting of many neurons (processing units) interconnected to form a complex network for expressing the complex number of uses between inputs and outputs in experimental data [41]. The multi-layer perceptron (MLP) is considered the most effective type of neural network, typically comprising a basic architecture of three layers. The trained data samples were fed through neurons into the input layer, the leftmost layer of the neural network. All nodes between the input and output layers constituted a hidden layer, which aided the neural network in learning the complex relationships between data, effectively serving as a layer for data processing. The neural network's final layer, derived from the first two layers, was the output layer, producing the final results. The number of neurons and hidden layers was determined by the complexity of the experimental data, utilizing known non-constant parameters to induce output variations, thereby handling many nonlinear data types. The model's inherent ability to provide nonlinear mappings between inputs and outputs is particularly beneficial for processing a significant amount of fuzzy or random data. The ANN algorithm in this study was implemented in a Python 3.10 environment, with functions used in the model added from the "standard scaler" function in the "sklearn" package.

Random Forest Regression Algorithm (RF)

The RF algorithm is one of the typical ensemble learning algorithms composed of decision trees [42]. The basic principle of the RF model consists of three main steps: random sampling, random selection of features, and majority voting. In the RF model, a subset was randomly selected from the calibration dataset, along with a random selection of some feature attributes. A decision tree model was then built using this subset and feature attributes. This process was repeated until the specified number of decision trees was established, which was then integrated into a random forest to obtain the final output result by averaging the predicted results. Compared to bagging, decision trees are randomly generated from a fixed-size subset of all attributes in RF, thereby reducing computational costs. RF has the advantage of high parallelization, significantly improving calibration speed on large data samples. Due to random sampling, the model exhibits strong generalization, low variance, high prediction accuracy, and good fitting. The RF algorithm in this study was implemented in a Python 3.10 environment, with the "Random Forest Regression" function from the "sklearn" package used during the computation process. The specific process is shown in Figure 4.

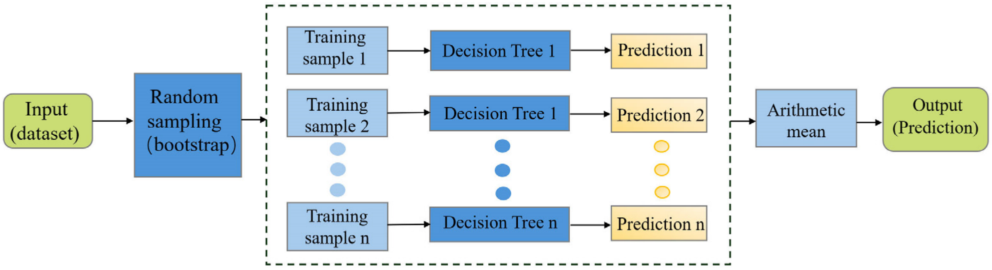


Figure 4. RF model flow chart.

2.3.4. Variable Importance Score

The relative importance of the target variable predicted by the RF model was evaluated through the relative depth of feature use. This was undertaken by assessing the contribution of each feature in the RF model across each tree, averaging these contributions, and comparing their magnitudes. The magnitude of contribution was typically based on the out-of-bag (OOB) error rate. Utilizing importance scores effectively assesses the contribution of each feature variable to the RF model, thereby reducing the complexity of model computation. The calculation formula for importance assessment is as follows:

$$\text{Importance score }(\text{x}) = \sum\nolimits_{\text{i}=1}^{\text{n}} \frac{\text{errOOB2} - -\text{errOOB1}}{\text{n}} \tag{2}$$

where n represents the number of decision trees, and errOOB1 and errOOB2 represent the error of the variable X in adding noise to a decision tree and the out-of-bag error.

2.3.5. Model Accuracy

Two methods, cross-validation and independent validation, were employed to validate the accuracy of the model. K-fold cross-validation involves dividing the original data into K subsets, using each subset once as a validation set, while the remaining K − 1 subsets are used as calibration sets. This ensures that each input variable is thoroughly analyzed and compared. In practical work, K must be sufficiently large to ensure an adequate number of calibration samples for each round. In this study, K was set to 10, which provided enough samples for cross-validation. The soil Cu regression model was validated using an independent validation dataset (Figure 1). Both validation methods were implemented in the Python 3.10 environment. These methods utilized the "Random Forest Regression" function from the "sklearn" package for calculation.

The performance of different spectral indices and models was evaluated by comparing the correlation coefficient (r), coefficient of determination (R^2), performance to deviation (RPD), relative error (RE, %), and standard error (RMSE). The closer R^2 and r values are to 1, the higher the RPD value, and the lower the RE and RMSE values, the better the accuracy and precision of the model. The formulas for calculating the above model evaluation parameters are as follows:

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

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

$$\mathrm{RPD} = \frac{\mathrm{SD}}{\mathrm{RMSE}} \quad (5)$$

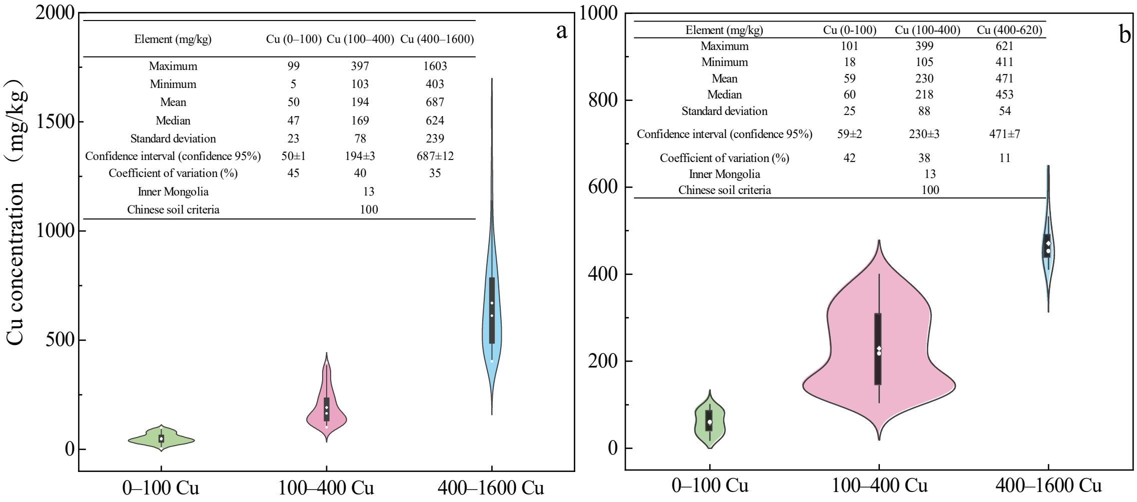
$$\mathrm{RE} = \frac{\mathrm{RMSE}}{\bar{y}_i} \times 100 \quad (6)$$

The meanings of all the letters in the formula above are as follows: n: number of samples; $\hat{y}_i$: measured values of Cu concentration; y_i: predicted value of Cu concentration; $\overline{y_i}$: average value of Cu concentration.

3. Results

3.1. Variation in Cu Concentration and Spectral Reflectance

The descriptive statistics of soil Cu concentrations obtained from laboratory measurements are depicted in Figure 5. Figure 5a illustrates that the Cu concentrations in soil samples from the Cu tailings area mainly ranged from 5 to 1603 mg/kg. Among all the soil samples obtained, 42% exhibited Cu concentrations exceeding the national pollution threshold in China [43]. A total of 72% of the samples surpassed the pollution background level [44]. In Figure 5b, it is shown that the Cu concentrations in the independent tailings area primarily fell within the range of 18–621 mg/kg. In this region, 16% of the soil samples exceeded the national pollution threshold in China. These results indicate a concentration of Cu elements within the study area.



Element (mg/kg)	Cu (0–100)	Cu (100–400)	Cu (400–1600)
Maximum	99	397	1603
Minimum	5	103	403
Mean	50	194	687
Median	47	169	624
Standard deviation	23	78	239
Confidence interval (confidence 95%)	50±1	194±3	687±12
Coefficient of variation (%)	45	40	35
Inner Mongolia		13	
Chinese soil criteria		100	

Element (mg/kg)	Cu (0-100)	Cu (100-400)	Cu (400-620)
Maximum	101	399	621
Minimum	18	105	411
Mean	59	230	471
Median	60	218	453
Standard deviation	25	88	54
Confidence interval (confidence 95%)	59±2	230±3	471±7
Coefficient of variation (%)	42	38	11
Inner Mongolia		13	
Chinese soil criteria		100	

Figure 5. Descriptive statistics of heavy metal concentrations include (**a**) Cu tailings mining area and (**b**) independent verification of the tailings area.

The original reflectance of soil Cu is shown in Figure 6. In the average scope, the standard deviation range is a partially transparent color. The spectral reflectance is negatively correlated with Cu concentration, with lower reflectance corresponding to higher Cu concentrations. Soil spectral calibration data and validation data are closely matched. A more comprehensive range of calibration datasets means enough generalization to ensure the model is universally applicable.

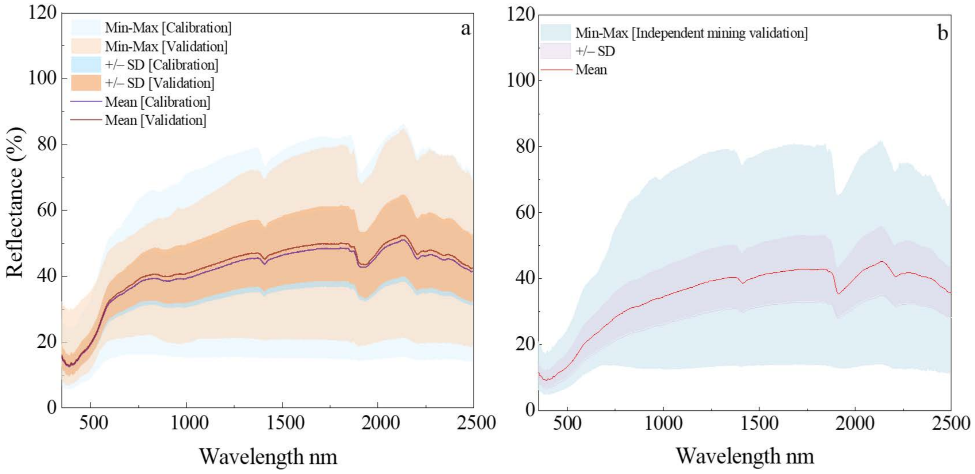


Figure 6. Spectral data of soil samples: (**a**) Cu field spectra and (**b**) independent tailings field spectra.

3.2. Relationship of Soil Cu Concentration against Spectral Indices

Figure 7 displays the optimized two-band contour map, while Figure 8 shows the three-dimensional slice plot of the optimized three-band spectral indices. Based on the highlighted spectral regions in Figures 7 and 8, the highest R^2 relationship between the optimal Opt-TBIs and Cu concentration is determined. Figure 9 shows the influence of Opt-TBIs on soil Cu concentration ranges from 41% to 59%. The near-infrared band (DIR, 750–1050 nm) is the primary sensitive band for the optimized two-band spectral indices in estimating soil Cu (Figure 10b). For the optimized three-band spectral indices, the sensitive bands are primarily located in the near-infrared (NIR, 750–1150 nm) region, with a small portion in the red edge (RE, 690–750 nm) region (Figure 10c).

The linear regression between the best-performing Opt-TBIs and Cu concentration is illustrated in Figures 11 and 12. The optimization algorithm for these bands significantly improves their sensitivity. Opt-TBI9 $(R_{1000} - R_{550})/(R_{1000} + R_{550} - 2 \times R_{1125})$ and Opt-TBI$_{12}$ $(R_{1150} + R_{700})/R_{975}$ exhibit the best performance among the spectral indices, with an R^2 of 0.59. Among all optimized spectral indices, the estimation capability of three-band spectral indices is higher than that of two-band spectral indices.

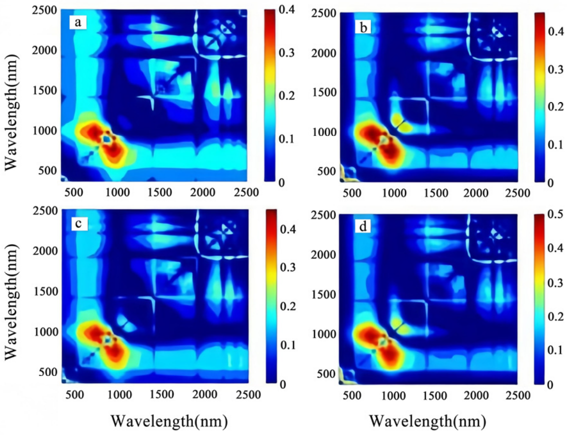


Figure 7. The examples of correlation of two-band indices and concentration of Cu: Opt-TBI$_1$ (**a**), Opt-TBI$_2$ (**b**), Opt-TBI$_3$ (**c**), Opt-TBI$_4$ (**d**).

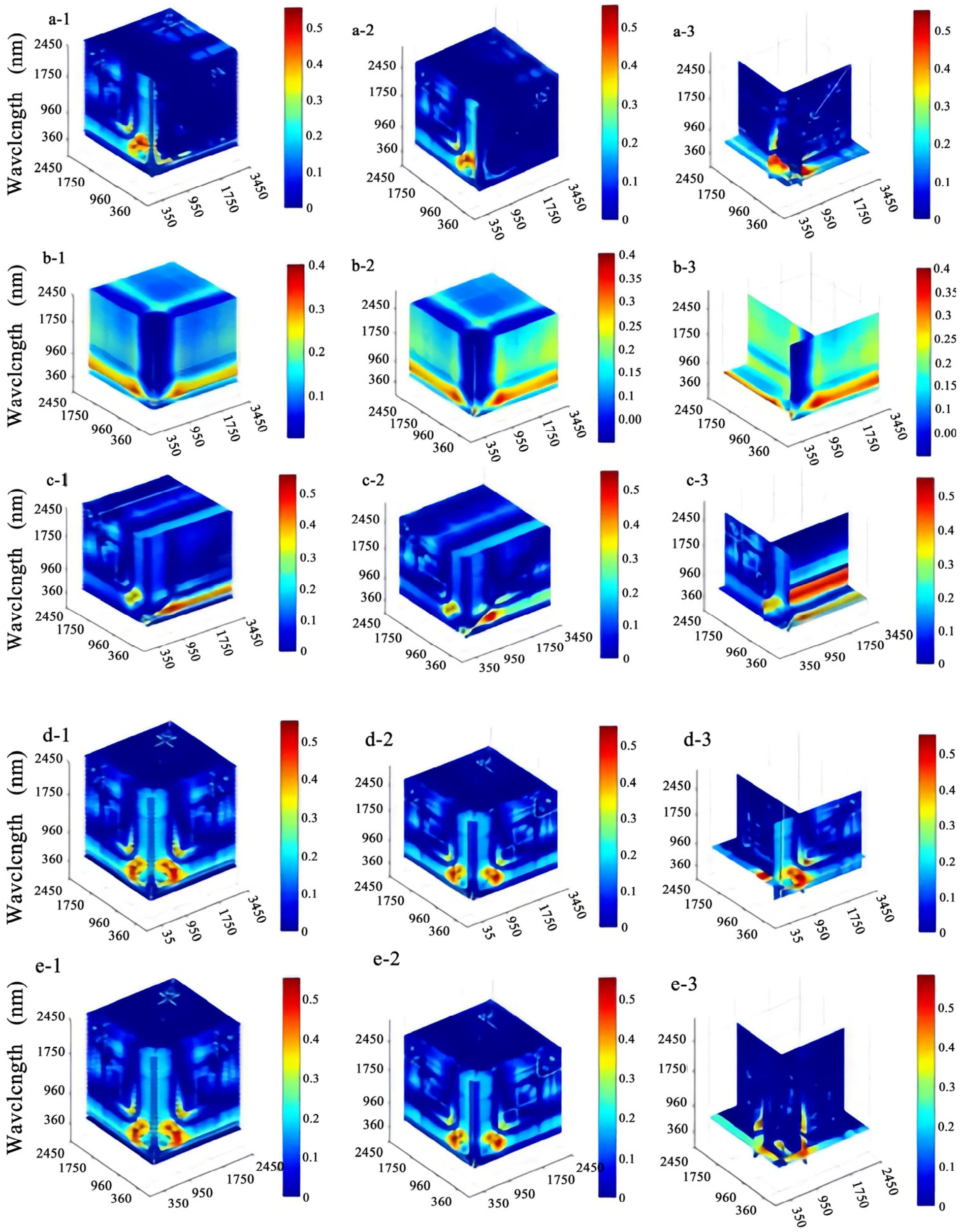
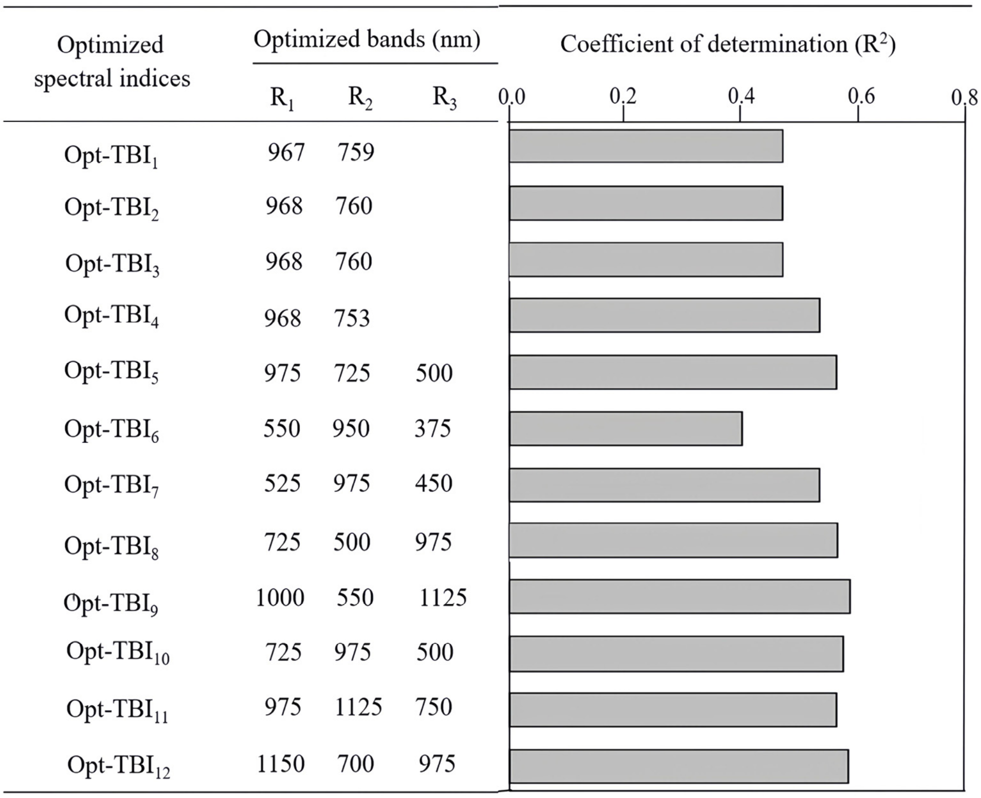


Figure 8. Soil Cu concentration and single three-band optimized spectral indices Opt-TBI$_5$ (**a**), Opt-TBI$_6$ (**b**), and Opt-TBI$_7$ (**c**). The R^2 slice of the relationship between Opt-TBI$_8$ (**d**) and Opt-TBI$_{11}$ (**e**) for all possible three-band combinations with a bandwidth of 1 nm in the growth stage range of 350–2500 nm (1: horizontal slice, 2: vertical slice, 3: the best film).



Optimized spectral indices	Optimized bands (nm)		
	R_1	R_2	R_3
$Opt\text{-}TBI_1$	967	759	
$Opt\text{-}TBI_2$	968	760	
$Opt\text{-}TBI_3$	968	760	
$Opt\text{-}TBI_4$	968	753	
$Opt\text{-}TBI_5$	975	725	500
$Opt\text{-}TBI_6$	550	950	375
$Opt\text{-}TBI_7$	525	975	450
$Opt\text{-}TBI_8$	725	500	975
$Opt\text{-}TBI_9$	1000	550	1125
$Opt\text{-}TBI_{10}$	725	975	500
$Opt\text{-}TBI_{11}$	975	1125	750
$Opt\text{-}TBI_{12}$	1150	700	975

Figure 9. The optimal band and the relationship between the optimized spectral indices and soil Cu concentration.

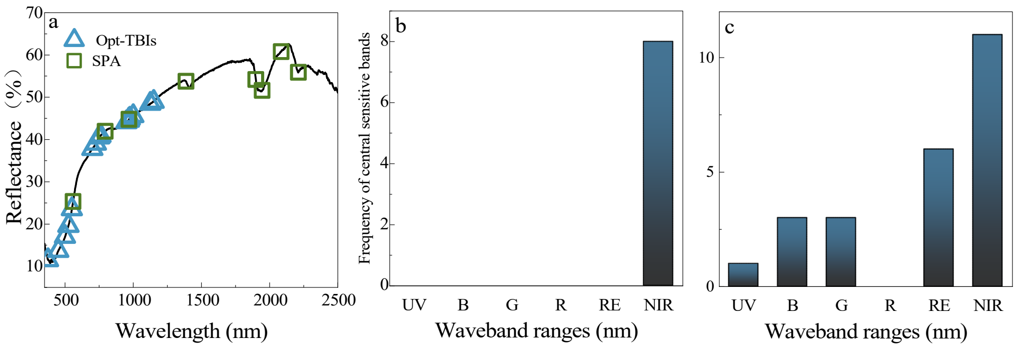


Figure 10. SPA and Opt-TBIs extract sensitive wavelength position (**a**), and two-band Opt-TBI (**b**) and three-band Opt-TBI sensitive band frequencies (**c**). Opt-TBIs (Ultraviolet radiation (UV): 340–400 nm; blue light (B): 450–520 nm, green light (G): 520–600 nm, red light range (R): 600–690 nm, red edge radiation (RE): 690–750 nm, near-infrared radiation range (NIR): 750–1150 nm).

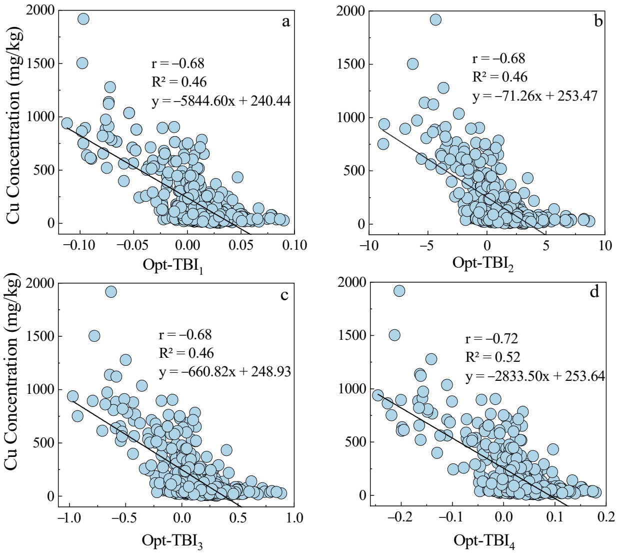


Figure 11. Correlation coefficients between two-band Opt-TBI$_1$ (**a**), Opt-TBI$_2$ (**b**), Opt-TBI$_3$ (**c**), and Opt-TBI$_4$ (**d**) based on soil spectral data and Cu concentration.

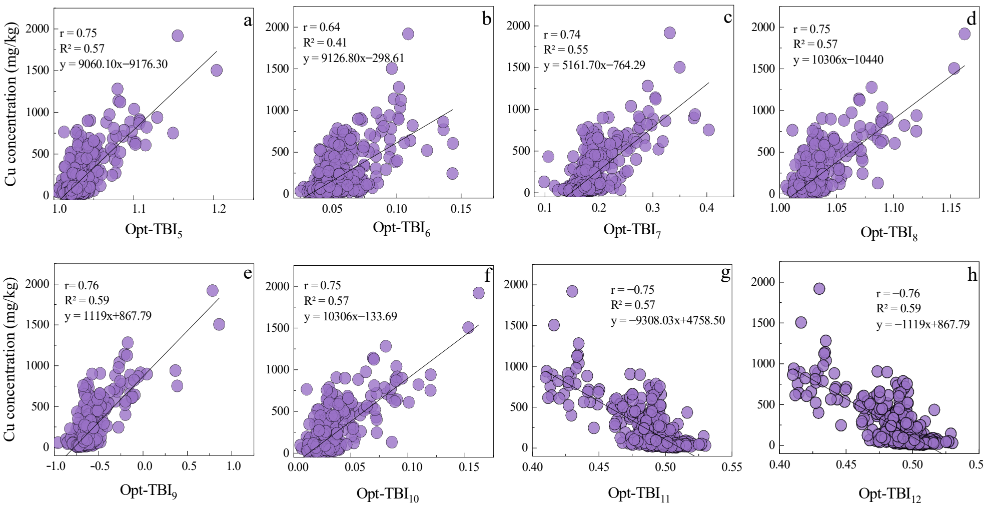


Figure 12. Correlation coefficients between TBI-band Opt-TBI$_5$ (**a**), Opt-TBI$_6$ (**b**), Opt-TBI$_7$ (**c**), Opt-TBI$_8$ (**d**) Opt-TBI$_9$ (**e**), Opt-TBI$_{10}$ (**f**), Opt-TB$_{11}$ (**g**), and Opt-TBI$_{12}$ (**h**) based on soil spectral data and Cu concentration.

3.3. Estimation of Cu Concentration Using a Machine Learning Model

Cu estimation performance analysis based on the machine learning model was performed on calibration datasets with different spectral input variables (Figure 13). The results show that the accuracy of the Cu concentration inversion model based on the original spectrum is about 0.5–0.9. With SPA optimization, the number of input variables can be significantly reduced, and the model's accuracy can be maintained. Compared with the original spectrum and the SPA, using the Opt-TBIs as the model input variable significantly improves the prediction ability of soil Cu in different calibration datasets, and the model inversion accuracy ranged from 0.67 to 0.95. The RF algorithm has the best prediction performance among the four models, followed by ANN and SVR. The best model for estimating soil Cu was Opt-TBIs-RF, with an R^2 of 0.95, RPD of 2.58, and RMSE of 61.68 mg/kg.

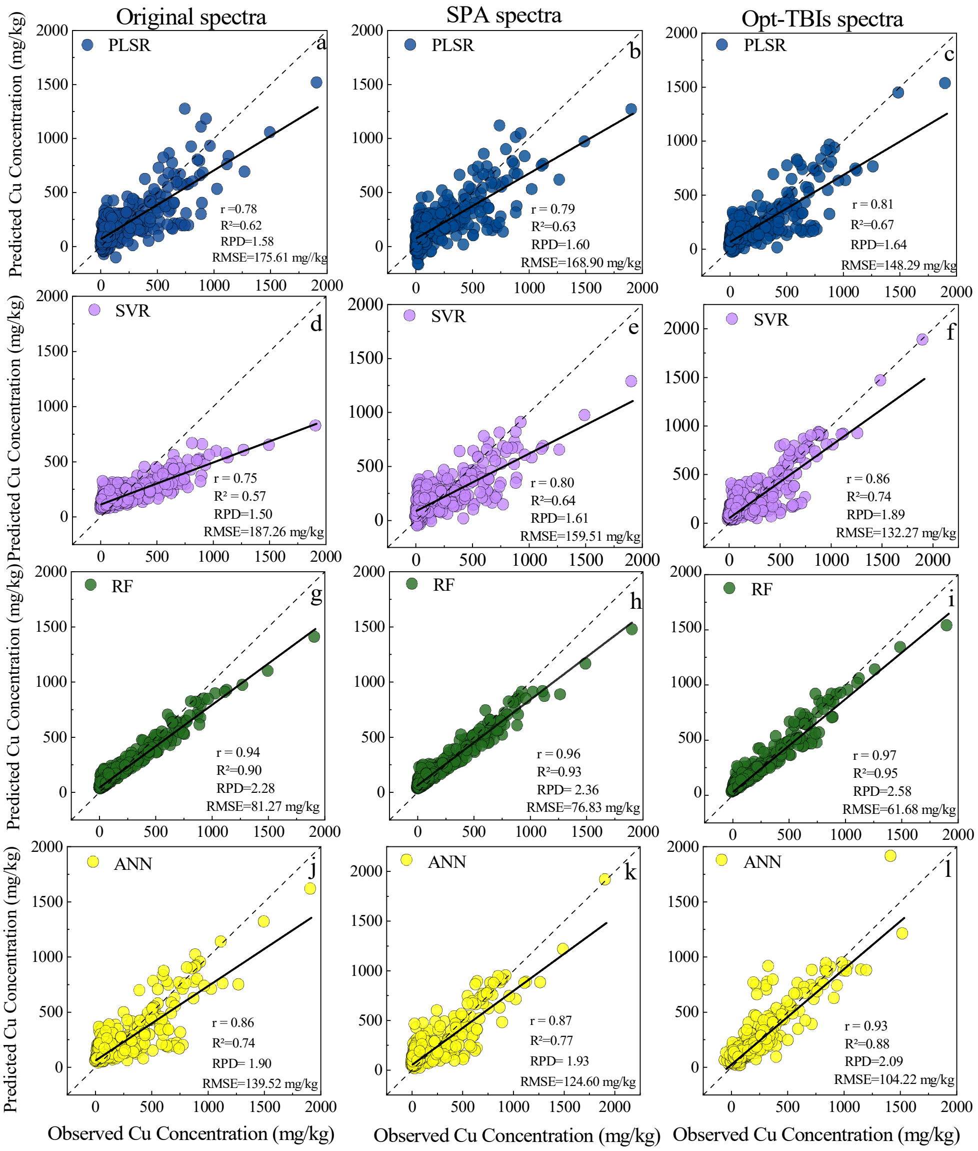


Figure 13. Cu concentration calibration set predictions using four models: full-band (**a**,**d**,**g**,**j**), characteristic band (**b**,**e**,**h**,**k**), and Opt-TBIs (**c**,**f**,**i**,**l**).

To better assess the performance of machine learning models in predicting soil Cu concentrations by combining different input variables, the accuracy of the established models was validated using a validation dataset (Figure 14). The results showed that the Opt-TBI combination with RF demonstrates the highest performance in predicting soil Cu concentrations, with an RPD of 2.31 and R^2 of 0.92. Compared to Opt-TBIs, raw spectra and the SPA exhibit poorer performance across different models, especially when the raw spectra are affected by redundant information.

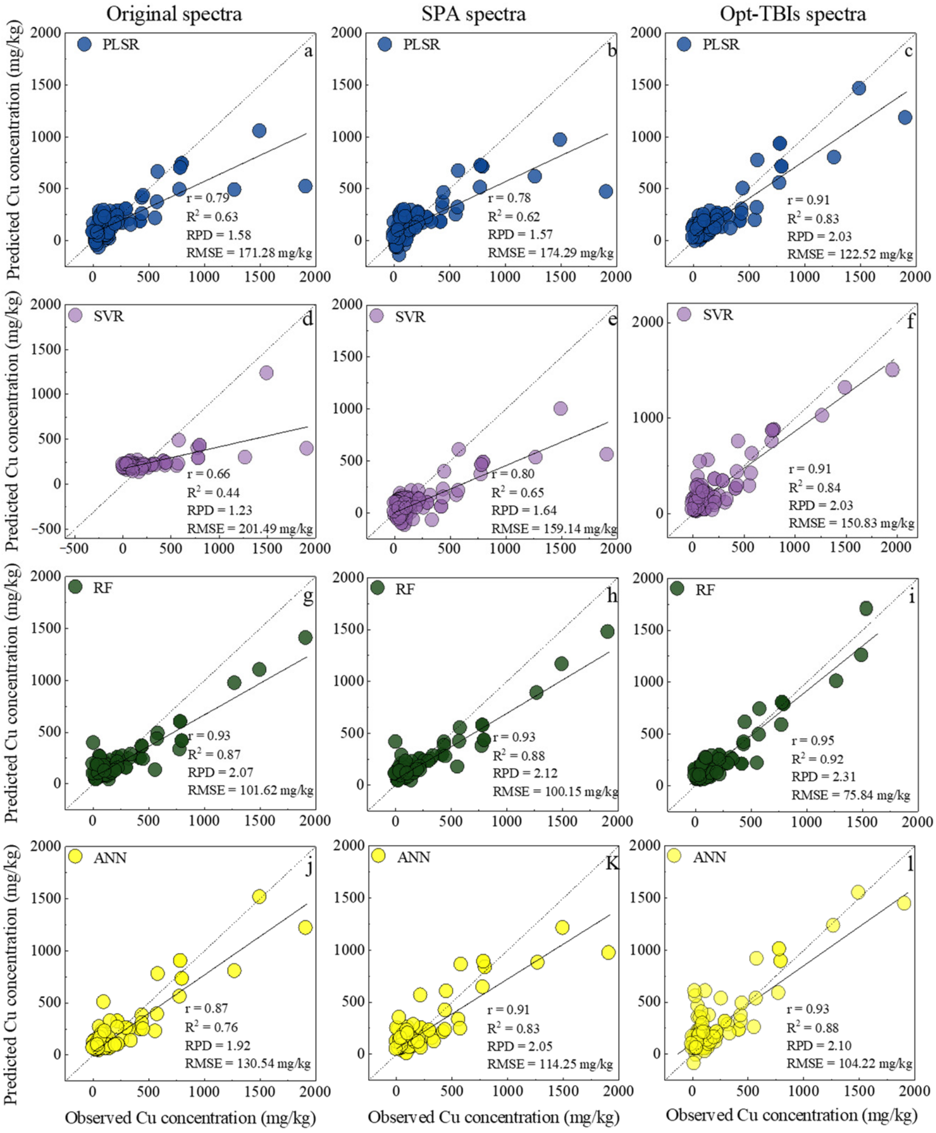


Figure 14. Cu concentration validation set predictions using four models: full-band (**a**,**d**,**g**,**j**), characteristic band (**b**,**e**,**h**,**k**), and Opt-TBIs (**c**,**f**,**i**,**l**).

3.4. Optimizing the Opt-TBIs-RF Model for Estimating Cu Concentration

To reduce the number of model input variables and obtain the most effective estimation model using fewer soil Cu characteristic bands, this study utilized the importance score of predictor variables predicted by the RF model to select the best input variables for the prediction model (Figure 15). Further analysis was conducted to examine the ranking of Opt-TBIs' contribution in the optimal RF model. The results revealed variations in the importance scores of the 12 Opt-TBIs (Figure 16). Opt-TBIs based on three bands have a relatively high importance score compared with those based on two bands, and, according to the ranking of model importance, Opt-TBI$_{10}$, Opt-TBI$_3$, Opt-TBI$_{11}$, and Opt-TBI$_2$ contributed the most to the RF model (Figure 16). The best Cu inversion model can be obtained using only the top four optimized spectral indices as input variables combined with the RF algorithm (Figure 17). Compared to establishing an RF model with 12 Opt-TBIs, incorporating indices with higher importance scores into the model improved computational efficiency and reduced the risk of overfitting. Utilizing the importance ranking of Opt-TBIs reduced the model's input variables by 33% and enhanced the stability of predicting soil Cu concentration.

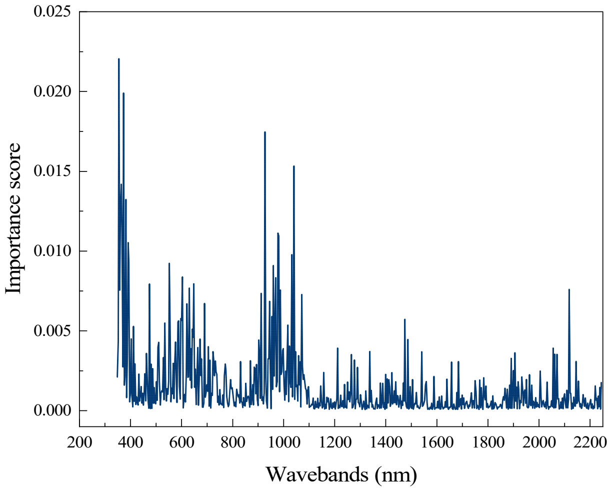


Figure 15. Comparison of full-spectrum-based RF models for importance scores in soil Cu.

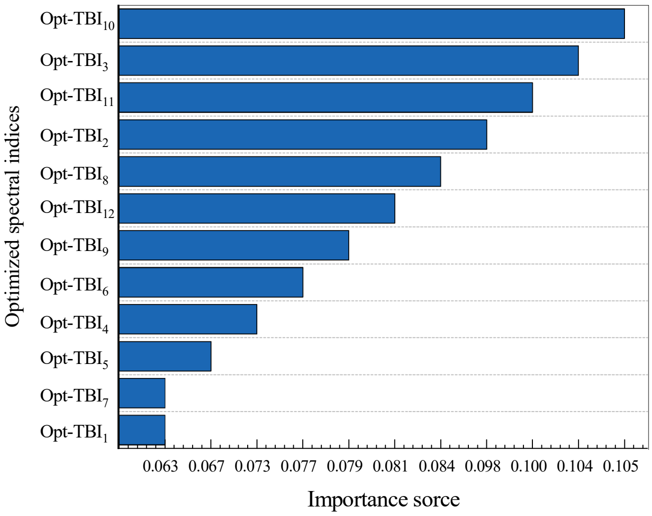


Figure 16. Comparison of importance scores of 12 Opt-TBI-based RF models in soil Cu.

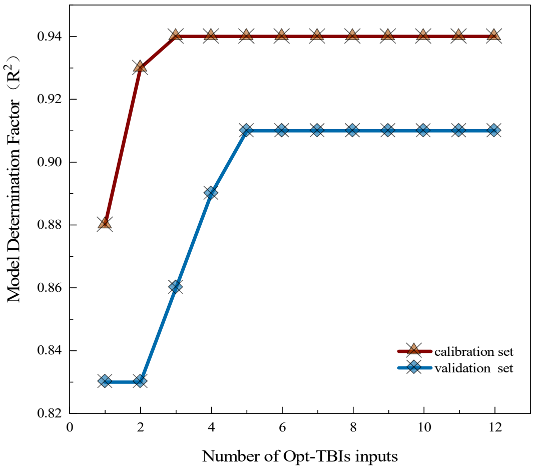


Figure 17. The number of spectra in the validation and calibration datasets was optimized based on the importance ranking of the Opt-TBIs and the RF model combination, and the performance of the model was stable when the first four Opt-TBIs were utilized.

The model dataset was cross- and independently validated to determine whether the modeling combination has a stable forecasting ability. The results in Figure 18 show that the R^2 and RMSE of the validation set of the model established by cross-validation are 0.92 and 74.78 mg/kg, respectively. Cross-validation fully uses limited datasets to better evaluate the model's performance under different data distributions. In contrast, the results of the independent validation model were also acceptable (R^2 = 0.91, RMSE = 77.21 mg/kg). The R^2 values of the validation methods were above 0.90, and the RPD values were above 2.0, indicating that the model had good stability for estimating soil Cu concentration.

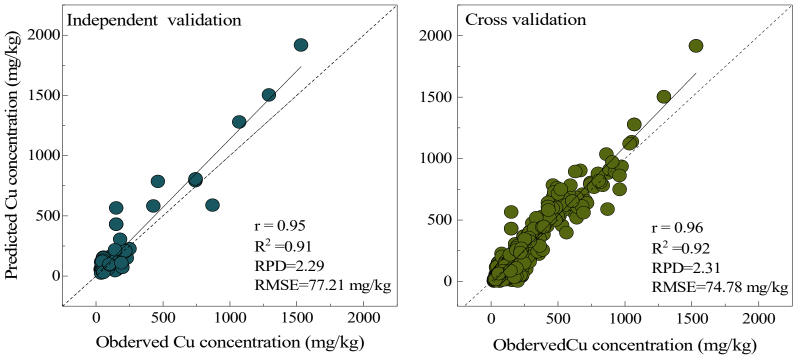


Figure 18. The accuracy of the Opt-TBIs-RF model was verified using independent and cross-datasets.

Using independent soil data from the tailings pond area as the model validation set, the Opt-TBIs-RF model's ability to predict soil Cu concentrations was verified for its generalization and accuracy. The results shown in Figure 19 indicate that the Opt-TBIs-RF model established using data from the independent tailings pond area can explain 72% of the variation in Cu concentrations in the soil. The model's RPD is 1.90, and the RMSE is 62.68 mg/kg. The estimation model constructed by Opt-TBIs-RF demonstrates an excellent linear relationship with the independent tailings pond area data, with the fitted data mainly distributed around the 1:1 line, indicating minimal deviation of the model.

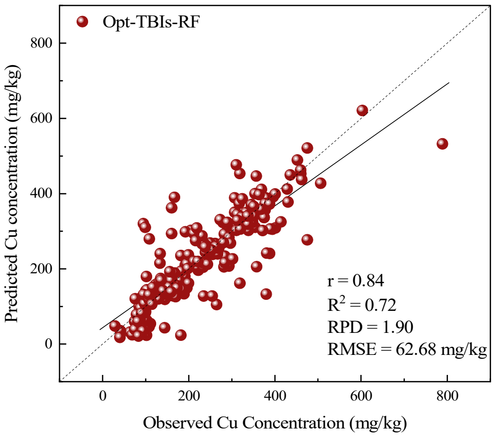


Figure 19. Verifies the generalization ability of the Opt-TBIs-RF model using an independent dataset from the tailings pond area.

4. Discussion

4.1. Comparison of Sensitive Wavebands

Extracting sensitive spectral bands from numerous soil spectra is crucial for improving the accuracy of predicting Cu concentration. Two feature selection methods, the SPA and Opt-TBIs, were employed to compare the extracted sensitive bands for soil Cu within the entire spectral range. The selected positions using the SPA and Opt-TBIs across the spectrum are illustrated in Figure 10a. It can be observed that some feature wavelengths selected through the two different methods were partially similar (e.g., 794 nm and 968 nm selected by the SPA, and 968 nm selected by Opt-TBIs). However, the SPA tends to select bands more concentrated in the near-infrared range, specifically around 1899–2211 nm. There are trace amounts of Cu pollutants formed by aluminosilicate minerals in the studied area. Previous studies have shown that the wavelength of Al-OH minerals mainly falls within the range of 1800–2200 nm [45]. However, since Cu compounds in this mining area are primarily distributed near oxides, the sensitive bands selected by SPA may have lower Cu information wavelengths to some extent. In contrast, Opt-TBIs select feature wavelengths with a stronger representativeness and specificity for soil Cu concentration. The soil Cu-sensitive bands identified by Opt-TBIs are predominantly located in the red edge (RE, 690–750 nm) and near-infrared (NIR, 750–1150 nm) regions, as depicted in Figure 10b,c. Since the mineral content in the mining area is mostly composed of hematite and magnetite, many pollution compounds will be formed with soil Cu. Previous studies have shown that wavelengths near 750 nm are related to iron and manganese oxides in the soil, where the active sites on the surfaces of iron and manganese oxides adsorb ions [46,47]. These active sites can further adsorb free Cu ions, highlighting more information about Cu in the soil [48,49]. In the sensitive bands extracted in this study, the near-infrared spectral range of 950–1150 nm contains more soil Cu-related information. This deviates from the conventional notion that the high-information zone for soil Cu is around 1200 nm in the organic matter spectral region [50–52]. This deviation may be due to the higher content of mineral elements in the soil of the mining area than the organic matter content. In soils containing a large amount of magnetite, metal ions are more likely to form complexes with OH- ions in magnetite [53]. Moreover, with increased Cu ion concentration, molecular groups in the near-infrared region (950–1150 nm) are more likely to form additional coordination compounds with Cu ions. Consequently, when Cu ions accumulate in the 950–1150 nm region, the extracted sensitive bands respond more to soil Cu concentration. Considering the spectral characteristics of soil Cu, this study accurately and precisely extracted bands containing high soil Cu information across the hyperspectral range (RE, 690–750 nm and NIR, 750–1150 nm). Future research can further optimize the sensitive bands for soil Cu by combining more characteristics of soil components.

4.2. Effects of Input Variables on Machine Learning Model Performance

Four widely used machine learning algorithms with different input variables were investigated to estimate the soil Cu concentration in the current study. Using soil Cu hyperspectral reflectance data, we compared the impact of different input variables on model performance (full-band, SPA, and Opt-TBI). The type and quantity of input variables significantly influenced the accuracy of machine learning models in estimating soil Cu concentration. The R^2 for the relationships between the full bands and Cu concentration in the four models in the validation and calibration dataset was relatively poor due to the influence of a large amount of redundant band information. The SPA can effectively eliminate the insensitive wavelength, reducing the model's complexity and calculation dimension. Similarly, Jia et al. [54] also confirmed that the PLSR model based on the effective wavelength of the SPA was significantly better in predicting soil nitrogen content than the PLSR model based on the whole band. Compared with the SPA, the Opt-TBI method is more agile and robust in deriving soil Cu concentration. A distinct aspect of this study is that Opt-TBIs are constructed by recombining soil Cu full spectrum bands rather than utilizing original formula bands. Results from Nawar et al. suggest that the SPA cannot accurately separate Cu spectral signals due to the highly overlapping nature of soil spectral information [55]. In contrast, Opt-TBIs, through a flexible combination of the full spectrum, better address the spectral overlap effect in soil, enhancing sensitivity to Cu concentration [56]. Compared to raw spectral indices, spectral indices exhibit improved predictive performance for soil Cu concentration by allocating more optimal band combinations. Opt-TBIs capture subtle spectral features of Cu concentration in soil, providing crucial input features for machine learning models. Integrating machine learning algorithms with sensitive bands extracted by Opt-TBIs reduces interference from irrelevant information and enhances the accuracy of soil Cu prediction [57,58]. Hence, machine learning algorithms based on Opt-TBIs represent a promising approach for predicting soil Cu concentration. Reducing input variables through Opt-TBIs requires fewer soil spectral features to accurately estimate soil Cu concentration. This effectively reduces the cost of monitoring soil Cu pollution, promoting more sustainable land management practices. In the future, rapid identification of regions with higher soil Cu concentrations can enable targeted development and implementation of soil conservation strategies to mitigate the impact of soil pollution on the environment and human health.

4.3. The Evaluation of Models

Before applying the model in practical applications, it is essential to validate its performance in inversely estimating soil Cu concentrations. In this study, to better assess the model's accuracy, we employed a cross-validation method to verify the Opt-TBIs-RF model on different subsets of soil Cu data. The model demonstrated excellent explanatory power on the cross-validation dataset, explaining 92% of the Cu concentration variations in the soil. However, previous studies have observed that cross-validation often utilizes internal data used in model development, resulting in deterministic model estimation outcomes [59,60]. To better mitigate this limitation, we introduced independent data from the surrounding study area for a second evaluation of the model's accuracy. Even under the 10% independent data validation scenario, the Opt-TBIs-RF model still achieved satisfactory estimation accuracy. However, validation using data from a single region alone cannot demonstrate the superior generalization ability of the model. In this study, independent soil Cu concentration data from tailings pond areas were obtained as the validation set to assess the adaptability of the Opt-TBI model to different datasets in diverse soil environments. The results demonstrated that the Opt-TBIs-RF model successfully predicted Cu concentrations in the independent tailings pond area soil data, showcasing the model's high generalization ability and feasibility. These research findings strongly attest that, when faced with appropriately expanded datasets, the Opt-TBIs-RF model exhibits robust adaptability, ensuring accurate and stable estimation of soil Cu concentrations.

To further validate the impact of sample concentration variability on model stability, soil Cu was categorized into three concentration gradients, and the RE values of the Opt-TBIs-RF model were compared at different Cu concentrations (Figures 20 and 21). It is well known that the maximum concentration in the calibration dataset will reduce the accuracy of the model's inversion of the target data [61]. However, in this study, the Opt-TBIs-RF model demonstrated stable estimation accuracy at the highest (400–1600 mg/kg) and lowest (0–100 mg/kg) concentrations of soil Cu. Remarkably, the Opt-TBIs-RF model maintained a robust linear relationship between predicted and laboratory-measured values when predicting concentrations of 100–400 mg/kg (Figure 20). This outcome effectively validates that using a calibration dataset spanning a wide range of spectral variations reduces the interference of sample concentration differences on the Opt-TBIs-RF model's estimation performance, thereby enhancing the model's predictive capability. Therefore, the Opt-TBIs-RF model emerges as a potentially reliable method for estimating soil Cu concentrations.

The soil Cu contamination can be rapidly and cost-effectively assessed through the Opt-TBIs-RF model, and targeted soil remediation and improvement measures can be implemented. For example, when elevated Cu levels are detected in soil, phytoremediation methods can be employed to reduce heavy metals accumulation. Additionally, early adoption of effective remediation measures can mitigate the soil Cu pollution to groundwater and surface water. Furthermore, monitoring and improving soil Cu concentrations can promote sustainable agricultural development, strengthen the quality of crop products, and foster ecological balance. The model will also be applied to monitor different soil environmental pollutants to further expand its practical value.

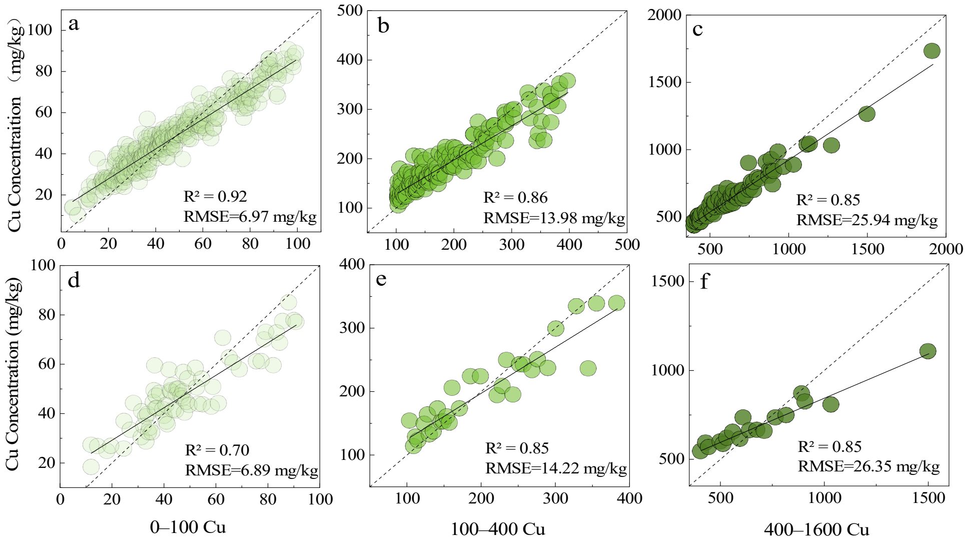


Figure 20. Correlation coefficients between the calibration set (**a**–**c**) and validation set (**d**–**f**) based on different Cu concentrations and RF models.

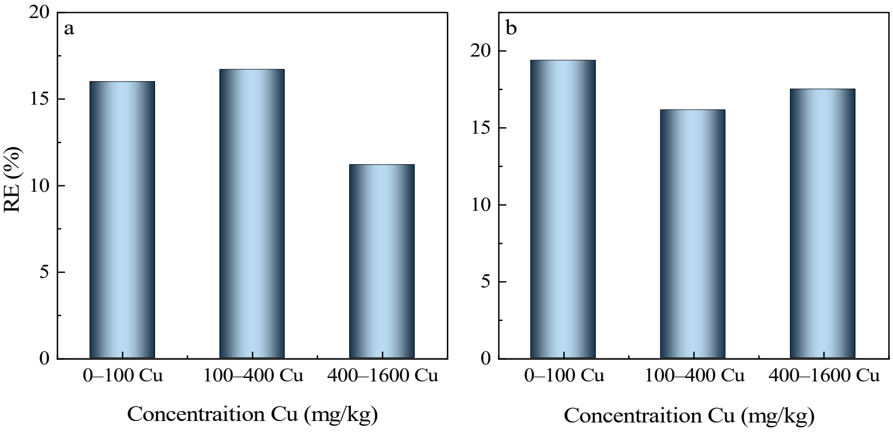


Figure 21. Effect of different soil Cu concentrations on the performance of the RF model. Model calibration set (**a**) and validation set (**b**).

5. Conclusions

In this study, soil Cu bands were optimized using full bands, the SPA, and Opt-TBIs, and thus, a quantitative relationship model for soil Cu concentration using the optimized spectral bands was established. The sensitive bands extracted by the SPA and Opt-TBI methods are closely associated with the characteristic bands of soil Cu concentration. Compared to spectral data extracted using full-band and SPA methods, the sensitive bands selected by Opt-TBIs effectively enhanced the estimation accuracy of the machine learning model. Utilizing Opt-TBIs combined with the RF model demonstrated the robustness in estimating soil Cu concentration. There were significant differences in the contribution of 12 optimized spectral indices to the RF model for assessing soil Cu concentration. Using the top four Opt-TBIs with the highest model importance as input variables, the RF model exhibited good stability and consistency across different mining area datasets and other Cu concentration distributions. In this study, the Opt-TBI algorithm showed higher predictive accuracy for soil Cu concentration in the research area. The predictive model based on Opt-TBIs can serve as a reference method for guiding the monitoring of soil Cu pollution in mining areas, with broad application prospects.

Author Contributions: Experiments were designed by M.H., F.L. and C.M. undertook the soil Cu concentration and hyperspectral band in the laboratory; C.M. compiled the data and performed the machine learning analysis; C.M. wrote the initial draft of the manuscript and F.L., C.M., M.H. and Y.H. edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: National Natural Science Foundation of China (M2042002), National key research and development plan project (2022YFD1500901), and Natural Science Foundation project of Inner Mongolia Autonomous Region (2021ZD10).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this study are available on request from the corresponding author.

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

References

1. Qiao, L.W.; Jiannan, C.; Feng, G.; Zi, J.L.; Meng, S.Z. Pollution level, ecological risk assessment and vertical distribution pattern analysis of heavy metals in the tailings dam of an abandon Lead–Zinc mine. *Sustainability* **2023**, *15*, 11987. [CrossRef]
2. Yong, P.G.; Xiao, J.L. Effects of bentonite addition on the speciation and mobility of Cu and Ni in soils from old mine tailings. *Sustainability* **2022**, *14*, 10878. [CrossRef]
3. Moses, A.A.; Belay, T.O.; Akash, K.; Jia, S.L.; Dayin, C.; Oluwaseun, P.O.; Lin, Z. Remediation of metal toxicity and alleviation of toxic metals-induced oxidative stress in *Brassica chinensis* L. using biochar-iron nanocomposites. *Plant Soil* **2023**, *493*, 629–645. [CrossRef]
4. Zhi, Y.L.; Zong, W.M.; Tsering, J.V.D.K.; Zeng, W.Y.; Lei, H. A review of soil heavy metal pollution from mines in China: Pollution and health risk assessment. *Sci. Total Environ.* **2014**, *468–469*, 843–853. [CrossRef] [PubMed]
5. Alazaiza, M.Y.D.; Albahnasawi, A.; Copty, N.K.; Bashir, M.; Nassani, D.; Maskari, A.T.; Amr, A.A.S.; Abujazar, M. Nanoscale zero-valent iron application for the treatment of soil, wastewater and groundwater contaminated with heavy metals: A review. *Desalin. Water Treat* **2022**, *253*, 194–210. [CrossRef]
6. Basegio, T.M.; Garcia, A.P.; Bergmann, C.P. Nanostructured zero-valent iron: From synthesis to application. *Environ. Appl. Nanomater.* **2022**, *152*, 205–237. [CrossRef] [PubMed]
7. Qiang, S.; Shi, W.Z.; Ke, X. Spectral heterogeneity analysis and soil organic matter inversion across differences in soil types and organic matter content in dryland farmland in China. *Sustainability* **2023**, *15*, 16310. [CrossRef]
8. Qing, Z.; Mamattursun, E.; Rukeya, S.; Mireguli, A.; Haoran, L.; Liling, W. Application of a hyperspectral remote sensing model for the inversion of nickel content in urban soil. *Sustainability* **2023**, *15*, 13948. [CrossRef]
9. Wen, X.G.; Yu, X.Z.; Jin, Y.X.; Ru, Q.Y.; Anna, X.; Xiao, D.H. Spatial distribution of soil heavy metal concentrations in road-neighboring areas using UAV-based hyperspectral remote sensing and GIS technology. *Sustainability* **2023**, *15*, 10043. [CrossRef]
10. Yun, X.; Bin, Z.; Yimin, W.; Yu, L.T.; Li, W.X. Hyperspectral inversion of chromium content in soil using support vector machine combined with lab and field spectra. *Sustainability* **2020**, *12*, 4441. [CrossRef]
11. Ming, S.Z.; Ying, F.G.; Yuan, Y.L.; Shi, H.W. Hyperspectral modeling of soil organic matter based on characteristic wavelength in east China. *Sustainability* **2022**, *14*, 8455. [CrossRef]
12. Bartholomeus, H.M.; Schaepman, M.E.; Kooistra, L.; Stevens, A.; Hoogmoed, W.B.; Spaargaren, O.S.P. Spectral reflectance-based indices for soil organic carbon quantification. *Geoderma* **2008**, *145*, 28–36. [CrossRef]
13. Mei, L.L.; Xiang, N.L.; Men, X.W.; Lu, F.L.; Lina, X. Integrating spectral indices with environmental parameters for estimating heavy metal concentrations in rice using a dynamic fuzzy neural-network model. *Comput. Geosci.* **2011**, *37*, 1642–1652. [CrossRef]
14. Wang, W.X.; Yao, X.F.; Yao, Y.C.; Tian, X.J.; Liu, J.; Ni, W.X.; Cao, Y.Z. Estimating leaf nitrogen concentration with three-band vegetation indices in rice and wheat. *Field Crops Res.* **2012**, *129*, 90–98. [CrossRef]
15. Shi, T.Z.; Wang, J.J.; Liu, H.Z.; Wu, G.F. Estimating leaf nitrogen concentrations in heterogeneous crop plants from hyperspectral reflectance. *Int. J. Remote Sens.* **2015**, *36*, 4652–4667. [CrossRef]
16. Datt, B. Visible/near infrared reflectance and chlorophyll content in Eucalyptus leaves. *Int. J. Remote Sens.* **2010**, *20*, 2741–2759. [CrossRef]
17. Venancio, L.P.; Mantovani, E.C.; Amaral, C.H.; Neale, C.M.U.; Gonçalves, I.Z.; Filgueiras, R.; Eugenio, F.C. Potential of using spectral vegetation indices for corn green biomass estimation based on their relationship with the photosynthetic vegetation sub-pixel fraction. *Agric. Water Manag.* **2020**, *236*, 106155. [CrossRef]
18. Zhu, Y.H.; Zhao, C.J.; Yang, H.; Yang, G.J.; Han, L.; Li, Z.H.; Feng, H.K.; Xu, B.; Wu, J.T.; Lei, L. Estimation of maize above-ground biomass based on stem-leaf separation strategy integrated with LiDAR and optical remote sensing data. *PeerJ* **2019**, *7*, e7593. [CrossRef] [PubMed]
19. Kooistra, L.; Salas, E.A.L.; Clevers, J.G.P.W.; Wehrens, R.; Leuven, R.S.E.W.; Niehuis, P.H.; Buydens, L.M.C. Exploring f-ield vegetation reflectance as an indicator of soil contamination in river floodplains. *Environ. Pollut.* **2004**, *127*, 281–290. [CrossRef]

20. Jiang, G.; Chen, X.; Wang, J.L.; Wang, S.S.; Zhou, S.U.; Bai, Y.; Liao, T.; Yang, H.; Ma, K.; Fan, X.L. Estimation of the multielement content in rocks based on a combination of visible–near-infrared reflectance spectroscopy and band index analysis. *Remote Sens.* **2023**, *15*, 3591. [CrossRef]
21. Xu, X.T.; Chen, S.B.; Ren, L.G.; Han, C.; Lv, D.L.; Zhang, Y.F.; Ai, F.K. Estimation of heavy metals in agricultural soils using Vis-NIR spectroscopy with fractional-order derivative and generalized regression neural network. *Remote Sens.* **2021**, *13*, 2718. [CrossRef]
22. Fu, P.J.; Yang, K.; Meng, F.; Zhang, W.; Cui, Y.; Feng, F.S.; Yao, G.B. A new three-band spectral and metal element index for estimating soil arsenic content around the mining area. *Process Saf. Environ.* **2022**, *157*, 27–36. [CrossRef]
23. Tan, K.; Wang, H.; Chen, L.H.; Du, Q.; Du, P.J.; Pan, C.C. Estimation of the spatial distribution of heavy metal in agricultural soils using airborne hyperspectral imaging and random forest. *J. Hazard. Mater.* **2020**, *382*, 120987. [CrossRef]
24. Cheng, H.; Shen, R.; Chen, Y.; Wan, Q.; Shi, T.; Wang, J.; Wan, Y.; Hong, Y.; Li, X. Estimating heavy metal concentrations in suburban soils with reflectance spectroscopy. *Geoderma* **2019**, *336*, 59–67. [CrossRef]
25. Shi, T.Z.; Chen, Y.L.; Liu, Y.L.; Wu, G.F. Visible and near-infrared reflectance spectroscopy—An alternative for monitoring soil contamination by heavy metals. *J. Hazard. Mater.* **2014**, *265*, 166–176. [CrossRef]
26. Wang, B.; Waters, C.; Orgill, S.; Cowie, A.; Clark, A.; Li, L.D.; Simpson, M.; McGowen, I.; Sides, T. Estimating soil organic carbon stocks using different modelling techniques in the semi-arid rangelands of eastern Australia. *Ecol. Indic.* **2018**, *88*, 425–438. [CrossRef]
27. Wang, S.; Chen, Y.H.; Wang, M.G.; Zhao, Y.F.; Li, J. SPA-based methods for the quantitative estimation of the soil salt content in saline-alkali land from field spectroscopy data: A case study from the yellow river irrigation regions. *Remote Sens.* **2019**, *11*, 967. [CrossRef]
28. Peng, X.T.; Shi, T.Z.; Song, A.H.; Chen, Y.Y.; Gao, W.X. Estimating soil organic carbon using VIS/NIR spectroscopy with SVMR and SPA methods. *Remote Sens.* **2014**, *6*, 2699–2717. [CrossRef]
29. Shi, R.J.; Pan, X.Z.; Wang, C.K.; Liu, Y.; Li, Y.L.; Li, Z.T. Prediction of cadmium content in the leaves of navel orange in heavy metal contaminated soil using VIS-NIR reflectance spectroscopy. *Spectrosc. Spectr. Anal.* **2015**, *35*, 3140–3145.
30. Han, C.; Lu, J.L.; Chen, S.B.; Xu, X.T.; Zi, B.W.; Zheng, P.; Yu, Z.; Feng, X.L. Estimation of heavy metal (Loid) contents in agricultural soil of the suzi river basin using optimal spectral indices. *Sustainability* **2021**, *13*, 12088. [CrossRef]
31. Wei, L.; Qiang, Y.; Teng, N.; Lin, Z.Y.; Hong, J.L. Inversion of soil heavy metal content based on spectral characteristics of peach trees. *Semant. Read.* **2021**, *12*, 1208. [CrossRef]
32. Shang, K.; Xiao, C.C.; Gan, F.P.; Wei, H.Y.; Wang, C.K. Estimation of soil copper content in mining area using ZY1-02D satellite hyperspectral data. *J. Appl. Remote Sens.* **2021**, *15*, 2607. [CrossRef]
33. An, R.; Wang, Y.X.; Zhang, X.W.; Chen, C.; Liu, X.Y.; Cai, S.T. Quantitative characterization of drying-induced cracks and permeability of granite residual soil using micron-sized X-ray computed tomography. *Sci. Total Environ.* **2023**, *876*, 163213. [CrossRef]
34. Jordan, C.F. Derivation of leaf-area index from quality of light on the forest floor. *Ecology.* **1969**, *50*, 663–666. [CrossRef]
35. Rouse, J.W.; Haas, R.H.; Deering, D.W.; Schell, J.A.; Harlan, J.C. *Monitoring the Vernal Advancement and Retrogradation (Green Wave Effect) of Natural Vegetation. [Great Plains Corridor]*; NASA: Washington, DC, USA, 1974.
36. Tucker, C.J. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sens. Environ.* **1979**, *8*, 127–150. [CrossRef]
37. Roujean, J.L.; Breon, F.M. Estimating PAR absorbed by vegetation from bidirectional reflectance measurements. *Remote Sens. Environ.* **1995**, *51*, 375–384. [CrossRef]
38. Gitelson, A.A.; Gritz, Y.; Merzlyak, M.N. Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. *J. Plant Physiol.* **2003**, *160*, 271–282. [CrossRef] [PubMed]
39. Penuelas, J.; Baret, F.; Filella, I. Semiempirical indices to assess carotenoids chlorophyll a ratio from leaf spectral reflectance. *Photosynthetica* **1995**, *31*, 221–230. [CrossRef]
40. Wang, F.H.; Gao, J.; Zha, Y. Hyperspectral sensing of heavy metals in soil and vegetation: Feasibility and challenges. *ISPRS J. Photogramm. Remote Sens.* **2018**, *136*, 73–84. [CrossRef]
41. García-Pedrajas, N.; Hervás-Martínez, C.; Muñoz-Pérez, J. COVNET: A cooperative coevolutionary model for evolving artificial neural networks. *IEEE Trans. Neural Netw.* **2003**, *14*, 575–596. [CrossRef]
42. Probst, P.; Wright, M.N.; Boulesteix, A.L. Hyperparameters and tuning strategies for random forest. *WIREs Data Min. Knowl. Discov.* **2019**, *9*, 1301. [CrossRef]
43. Sun, Y.; Li, H.; Guo, G.L.; Semple, K.T.; Jones, K.C. Soil contamination in China: Current priorities, defining background levels and standards for heavy metals. *J. Environ. Manag.* **2019**, *251*, 109512. [CrossRef]
44. Bo, Z.; Bin, G.; Bin, Z.; Wei, W.; Yong, Z.L.; Tianqi, L. Retrieving soil heavy metals concentrations based on GaoFen-5 hyperspectral satellite image at an opencast coal mine, Inner Mongolia, China. *Environ. Pollut.* **2022**, *24*, 118981. [CrossRef]
45. Clark, R.N. *Chapter 1: Spectroscopy of Rocks and Minerals, and Principles of Spectroscopy, in Manual of Remote Sensing, Volume 3, Remote Sensing for the Earth Sciences*; John Wiley and Sons: New York, NY, USA, 1999; pp. 3–58.
46. Bedini, E. The use of hyperspectral remote sensing for mineral exploration: A review. *J. Hyperspectral Remote Sens.* **2017**, *7*, 189–211. [CrossRef]

47. Jiang, G.; Zhou, S.G.; Cui, S.C.; Chen, T.; Wang, J.L.; Chen, X.; Liao, S.B.; Zhou, K. Exploring the potential of HySpex hyperspectral imagery for extraction of copper content. *Sensors* **2020**, *20*, 6325. [CrossRef]
48. Bendor, E.; Banin, A. Near-infrared analysis as a rapid method to simultaneously evaluate several soil properties. *Soil Sci. Soc. Am. J.* **1995**, *59*, 364–372. [CrossRef]
49. Sun, F.S.; Li, Y.Q.; Wang, X.; Chi, Z.L.; Yu, G.H. Using new hetero-spectral two- dimensional correlation analyses and synchrotron-radiation-based Spectro microscopy to characterize binding of Cu to soil dissolved organic matter. *Environ. Pollut.* **2017**, *223*, 457–465. [CrossRef]
50. Komy, Z.R.; Shaker, A.M.; Heggy, S.E.M.; El-Sayed, M.E.A. Kinetic study for copper adsorption onto soil minerals in the absence and presence of humic acid. *Chemosphere* **2014**, *99*, 117–124. [CrossRef] [PubMed]
51. Yin, F.; Wu, M.M.; Liu, L.; Zhu, Y.Q.; Feng, J.L.; Yin, D.W.; Yin, C.J.; Yin, C.T. Predicting the abundance of copper in soil using reflectance spectroscopy and GF5 hyperspectral imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *102*, 10242. [CrossRef]
52. Tu, Y.L.; Zou, B.; Jiang, X.L.; Tao, C.; Tang, Y.Q.; Feng, H.H. Hyperspectral remote sensing-based modeling of Cu content in mining soil. *Spectrosc. Spectr. Anal.* **2018**, *38*, 575–581.
53. Li, P.; Fan, Q.; Pan, D.; Liu, S.; Wu, W. Effects of pH, ionic strength, temperature, and humic acid on Eu (III) sorption onto iron oxides. *J. Radioanal. Nucl. Chem.* **2011**, *289*, 757–764. [CrossRef]
54. Jia, S.Y.; Zhang, J.M.; Li, G.; Yang, X.L. Predicting soil nitrogen and organic carbon using near infrared spectroscopy coupled with variable selection. *Appl. Eng. Agric.* **2014**, *30*, 641–647. [CrossRef]
55. Nawar, S.; Buddenbaum, H.; Hill, J. Estimation of soil salinity using three quantitative methods based on visible and near-infrared reflectance spectroscopy: A case study from Egypt. *Arab. J. Geosci.* **2015**, *8*, 5127–5140. [CrossRef]
56. Kooistra, L.; Wanders, J.; Epema, G.F.; Leuven, R.S.E.W.; Wehrens, R.; Buydens, L.M.C. The potential of field spectroscopy for the assessment of sediment properties in river floodplains. *Anal. Chim. Acta* **2003**, *484*, 189–200. [CrossRef]
57. Bao, N.; Wu, L.; Ye, B.; Yang, K.; Zhou, W. Assessing soil organic matter of reclaimed soil from a large surface coal mine using a field spectroradiometer in laboratory. *Geoderma* **2017**, *288*, 47–55. [CrossRef]
58. Shi, T.Z.; Liu, H.Z.; Chen, Y.Y.; Wang, J.J.; Wu, G.F. Estimation of arsenic in agricultural soils using hyperspectral vegetation indices of rice. *J. Hazard. Mater* **2016**, *308*, 243–252. [CrossRef] [PubMed]
59. Forkuor, G.; Hounkpatin, O.K.L.; Welp, G.; Thiel, M.; Hui, D.F. High resolution mapping of soil properties using remote sensing variables in south-western burkina faso: A comparison of machine learning and multiple linear regression models. *PLoS ONE* **2017**, *12*, 0170478. [CrossRef] [PubMed]
60. Rivera, J.L.; Bonilla, C.A. Predicting soil aggregate stability using readily available soil properties and machine learning techniques. *Catena* **2020**, *187*, 104408. [CrossRef]
61. Moura-Bueno, J.M.; Dalmolin, R.S.D.; Ten, C.A.; Dotto, A.C.; Demattê, A.M. Stratification of a local VIS-NIR-SWIR spectral library by homogeneity criteria yields more accurate soil organic carbon predictions. *Geoderma* **2019**, *337*, 565–581. [CrossRef]

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.