



Article

Sentinel-2 Multispectral Satellite Remote Sensing Retrieval of Soil Cu Content Changes at Different pH Levels

Hongxu Guo ¹, Fan Wu ¹ , Kai Yang ¹, Ziyang Yang ¹, Zeyu Chen ¹, Dongbin Chen ¹ and Rongbo Xiao ^{2,*} 

¹ School of Architecture and Urban Planning, Guangdong University of Technology, Guangzhou 510090, China; guohx@gdut.edu.cn (H.G.); 2112210049@mail2.gdut.edu.cn (F.W.); 2112210059@mail2.gdut.edu.cn (K.Y.); 2112210047@mail2.gdut.edu.cn (Z.Y.); 2112210056@mail2.gdut.edu.cn (Z.C.); 2112310039@mail2.gdut.edu.cn (D.C.)

² School of Geography and Environmental Economics, Guangdong University of Finance & Economics, Guangzhou 510320, China

* Correspondence: ecoxiaorb@163.com; Tel.: +86-1342-366-6257

Abstract: With the development of multispectral imaging technology, retrieving soil heavy metal content using multispectral remote sensing images has become possible. However, factors such as soil pH and spectral resolution affect the accuracy of model inversion, leading to low precision. In this study, 242 soil samples were collected from a typical area of the Pearl River Delta, and the Cu content in the soil was detected in the laboratory. Simultaneously, Sentinel-2 remote sensing image data were collected, and two-dimensional and three-dimensional spectral indices were established. Constructing independent decision trees based on pH values, using the Successive Projections Algorithm (SPA) combined with the Boruta algorithm to select the characteristic bands for soil Cu content, and this was combined with Optuna automatic hyperparameter optimization for ensemble learning models to establish a model for estimating Cu content in soil. The research results indicated that in the SPA combined with the Boruta feature selection algorithm, the characteristic spectral indices were mainly concentrated in the spectral transformation forms of TBI2 and TBI4. Full-sample modeling lacked predictive ability, but after classifying the samples based on soil pH value, the R^2 of the RF and XGBoost models constructed with the samples with pH values between 5.85 and 7.75 was 0.54 and 0.76, respectively, with corresponding RMSE values of 22.48 and 16.12 and RPD values of 1.51 and 2.11. This study shows that the inversion of soil Cu content under different pH conditions exhibits significant differences, and determining the optimal pH range can effectively improve inversion accuracy. This research provides a reference for further achieving the efficient and accurate remote sensing of heavy metal pollution in agricultural soil.

Keywords: Sentinel-2; spectral indices; decision tree; Optuna; ensemble learning



Citation: Guo, H.; Wu, F.; Yang, K.; Yang, Z.; Chen, Z.; Chen, D.; Xiao, R. Sentinel-2 Multispectral Satellite Remote Sensing Retrieval of Soil Cu Content Changes at Different pH Levels. *Agronomy* **2024**, *14*, 2182. <https://doi.org/10.3390/agronomy14102182>

Academic Editors: Yash Dang and Enrico Corrado Borgogno Mondino

Received: 24 July 2024

Revised: 13 September 2024

Accepted: 19 September 2024

Published: 24 September 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With the acceleration of industrialization, urbanization, and agricultural intensification, the problem of soil heavy metals is becoming increasingly prominent [1]. Heavy metal pollution mainly refers to soil contamination by heavy metals and their compounds resulting from processes such as fertilizer use, electronic waste, pesticides, herbicides, and industrial waste treatment. According to the nationwide soil pollution status survey conducted by the Ministry of Ecology and Environment of the People's Republic of China, the excessive rate of pollution points in arable land soil in China has reached 19.4%, with heavy metals being the main pollutants. Among them, the excessive rate of Cu pollution points has reached 2.1%. Copper (Cu) is an essential trace element for all living organisms, but it can become toxic if threshold concentrations maintained by evolutionarily conserved homeostatic mechanisms are exceeded [2]. Traditional monitoring methods for soil composition in agricultural land are often hindered by limitations in manpower, time, and geographical coverage, leading to inefficiency and high costs, and it is difficult to achieve

precise monitoring on a large scale. Remote sensing technology, with its high efficiency, can rapidly retrieve information on crops and soil composition in agricultural land, effectively addressing the shortcomings of traditional methods and providing strong support for soil protection efforts. Based on remote sensing images, the rapid monitoring of agricultural land can be achieved, thereby determining the areas and levels of soil heavy metal pollution that need remediation.

Numerous studies have applied multispectral satellite imagery to accurately retrieve key components in various crops and soils [3], such as crop chlorophyll content [4], biomass [5], soil salinity [6,7], organic carbon [8,9], nitrogen, phosphorus, potassium [3,10], and other important indicators [10,11]. Soil heavy metals are trace elements with weak spectral responses at medium to low concentrations and are readily adsorbed by organic matter, clay minerals, and other substances [12,13]. Consequently, the characteristic signals in the raw spectral data are often obscured, making the identification and extraction of representative feature variables from multispectral data complex and challenging. Previous studies have constructed spectral indices to explore more spectral information in multispectral data, which can provide sensitive spectral differentiation information for the inversion of soil heavy metals [14]. Yu [15] utilized the reflectance data of nine original bands from Sentinel-2A to estimate the content of Pb and Cd in soil. By employing the inverse ($1/B_i$), logarithmic ($\ln B_i$), and band ratio (B_i/B_j) transformation methods, along with the terrain features and spatial characteristics of pollution sources, the accuracy of the inversion model was significantly improved. Song [14] used principal component analysis (PCA) on the original bands of Landsat 7 ETM+ and Landsat 8 OLI images, as well as their respective band transformations and vegetation indices, to model and improve the spatial estimation accuracy of soil heavy metal content. Yang [16] utilized hyperspectral-simulated Landsat 8 OLI multispectral data to calculate key indices such as the normalized difference vegetation index (NDVI), the modified normalized difference water index (MNDWI), the difference vegetation index (DVI), the enhanced vegetation index (EVI), and the clay mineral ratio (CMR). Additionally, characteristic components such as greenness, brightness, and wetness were considered. By combining these factors with partial least squares regression (PLSR) modeling, the multispectral remote sensing mapping of Hg was accomplished. Despite a significant amount of work performed on constructing spectral indices using multispectral images for soil heavy metal inversion, the accuracy of inversion still needs improvement [17]. Furthermore, the currently used spectral indices are mostly common band variations or vegetation indices, and we have not comprehensively explored all possible two-dimensional and three-dimensional spectral indices.

Soil composition is complex and significantly influenced by external conditions, making spectral measurements prone to errors. Specifically, factors such as pH value, moisture content, and organic matter can all significantly affect spectral reflectance [18]. Due to the difference in soil pH value, Cu will present different physical and chemical states in the soil [19]. In acidic soil, Cu may exist as soluble ions. In neutral or alkaline soils, Cu is often combined with organic matter or oxide to form a more stable complex or precipitation. Soil pH indirectly affects the spectral reflectance by affecting the shape and color of Cu-containing compounds, and thus affects the inversion results of soil heavy metals [20–22]. In addition, soil pH strongly correlates with soil compaction, soil moisture, vegetation growth status, etc. Soil pH indirectly affects spectral reflectance, influencing soil heavy metal inversion results. As the alkalinity of the soil increases, the intensity of soil compaction increases, the soil surface becomes smoother, the brightness of the color also increases, and the reflectance of the soil changes accordingly [20]. Meanwhile, there is a significant negative correlation between soil moisture and pH value [23]. When the pH value decreases, the net charge of the soil increases, weakening the cohesion strength between soil particles, leading to soil structure loosening, and causing changes in soil reflectance. Furthermore, the nutritional growth and yield of most crops decrease significantly under lower pH conditions, while they tend to increase as the pH approaches the optimal level. Many crops exhibit optimal growth over a near-neutral pH range, while a

few crops thrive better in either acidic or alkaline soils [24]. Factors such as soil pH can obscure or alter the absorption characteristics of soil heavy metals in the spectra, thereby reducing the predictive accuracy of models. Previous studies typically focused only on the impact of soil moisture, employing linear methods such as Direct Standardization (DS), Piecewise Direct Standardization (PDS), and External Parameter Orthogonalization (EPO) to calibrate original soil samples using dried soil samples. This approach aims to eliminate or mitigate the effects of soil moisture on spectral reflectance [25,26]. However, when inverting large-scale soil heavy metal content, there are significant differences in the physicochemical properties of soil samples. Linear correction algorithms alone are insufficient to completely eliminate the impacts of physicochemical properties on spectral reflectance, leading to distortions in the inversion models. Even without considering the possibility of complete correction, it remains very difficult to correct the spectral reflectance of samples from one pH range to another specific pH range. This is due to the inability to determine the exact range where the spectral reflectance is least affected by pH, and the process of calibrating the samples' pH to a specific range is complex and cumbersome.

The aim of this study was to explore the feasibility of using multispectral data to invert the Cu content in soil. Focusing on the typical regions of the Pearl River Delta, Sentinel-2 multispectral data were used to construct common two-dimensional and three-dimensional spectral indices to extract more spectral information from the multispectral data and to investigate the optimal spectral transformation forms. A decision tree based on pH value was constructed to divide the samples and determine the optimal pH range for soil Cu content inversion, thereby enhancing the sensitivity of spectral indices to soil heavy metal Cu content. Meanwhile, Optuna was used to evaluate the impact of various hyperparameters on the RMSE loss function, and combined with ensemble learning algorithms, a high-precision inversion model for soil Cu content was established.

2. Materials and Methods

In this study, we developed a model based on Sentinel-2A imagery for the high-precision inversion of Cu content in soil. The workflow is shown in Figure 1. The study consisted of four main steps. (1) Data acquisition and pre-processing: In the laboratory, the heavy metal content of collected soil samples was measured, and the spectral reflectance of each soil sample in each band was extracted from the multispectral images. To amplify the effect of heavy metals on soil spectral reflectance characteristics, two-dimensional and three-dimensional spectral indices were constructed from the original bands. (2) Soil sample classification and feature selection: Using pH value as the branching criterion, a decision tree was constructed to classify the samples. Then, the Successive Projections Algorithm (SPA) combined with the Boruta algorithm was used to extract the characteristic variables of Cu from each sample set. (3) Cu content estimation model building: The bagging-strategy-based Random Forest (RF) and the boosting-strategy-based Extreme Gradient Boosting (XGBoost) models were constructed to estimate Cu content. (4) Comparison of inversion accuracy: The model's inversion accuracy was evaluated using three precision indicators: R^2 , RMSE, and RPD.

2.1. Study Area

The study area, encompassing 3330 square kilometers, is located in the western wing of the Pearl River Delta Greater Bay Area, with geographical coordinates ranging from $113^{\circ} 6' 10''$ to $114^{\circ} 19' 9''$ E and $21^{\circ} 50' 13.3''$ to $22^{\circ} 46' 34.26''$ N, as shown in Figure 2. The terrain of this area is flat, with a dense river network, and an average elevation of 19 m. The area has a subtropical monsoon climate, with a mild climate and abundant rainfall, and a maritime climate moderates it [27]. The annual average temperature is about 23°C , the annual average precipitation reaches 122 mm, and the average wind speed is 16 km/h. Red soil is the main soil type in this area, typically acidic, which results in some heavy metals existing in ionic form [28]. At the same concentration, heavy metals in acidic soils pose a higher potential risk of environmental pollution than in alkaline soils.

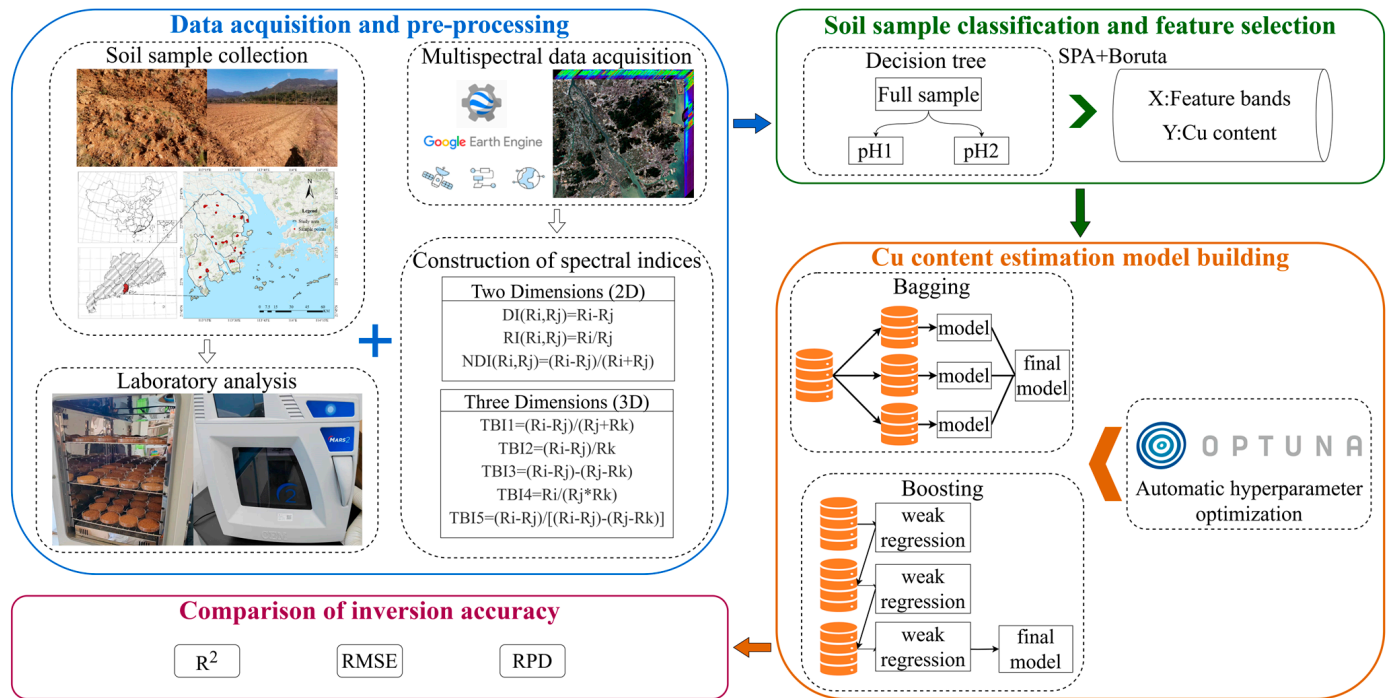


Figure 1. The flowchart of study.

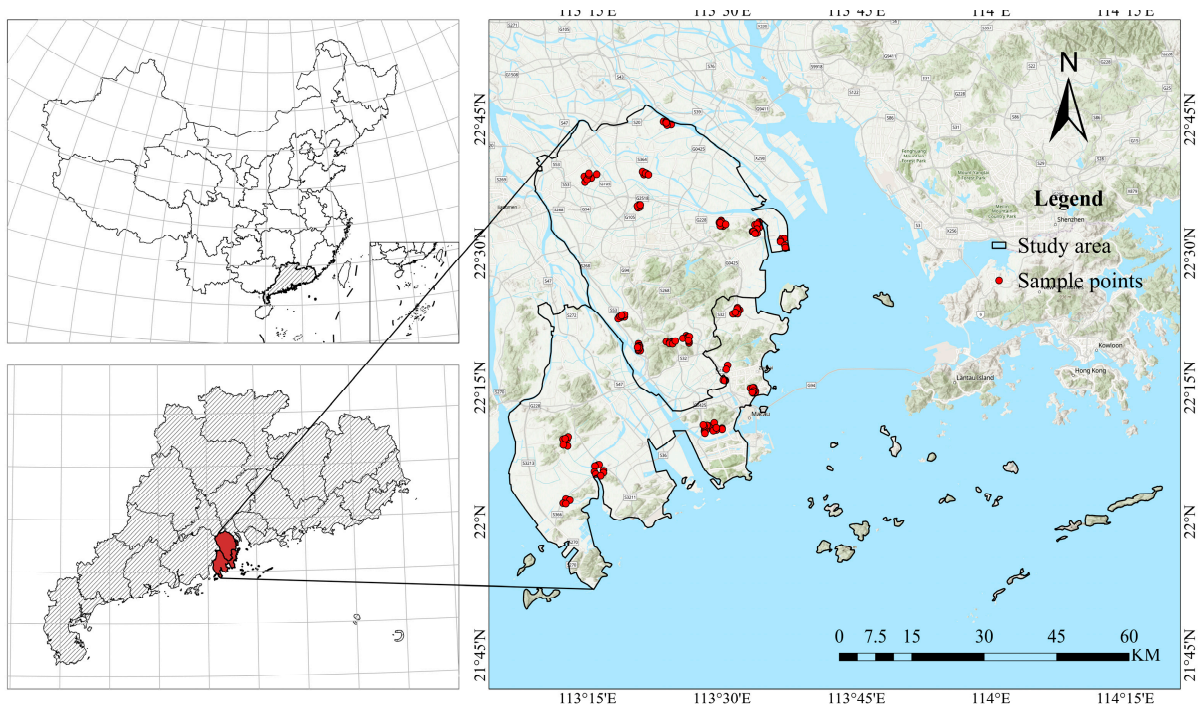


Figure 2. General situation of the research region and distribution of sampling points.

2.2. Soil Sample Collection and Laboratory Analysis

To ensure the scientific rigor and effectiveness of the investigation, we adopted a series of strict screening criteria when selecting the survey subjects:

- The enterprise’s land area should generally not be less than 7000 square meters;
- The main production facilities and the three-waste treatment facilities should generally not be located on the second floor or above;
- No soil-pollution-related enterprises should exist within a 500 m radius of the enterprise;
- Agricultural land should be present within a 3 km radius of the enterprise;

- The enterprise must have been in operation at its current location for at least five years;
- There should be no history of other soil-polluting enterprises at the enterprise's location;
- The number of other enterprises around the location should be relatively small.

Based on these criteria, we ultimately selected 20 typical soil pollution enterprises and collected soil samples from the farmland surrounding these enterprises. Additionally, considering factors such as farmland layout, topography, and road accessibility, a total of 242 sampling points were comprehensively selected, as shown in Figure 2. The study area's surface soil samples were collected from 15 June 2022 to 20 July 2022. This study employed a five-point sampling method, wherein five independent surface soil samples (0–20 cm) were collected at each sampling point and then mixed for preparation. Additionally, we utilized real-time kinematic (RTK) technology to record the latitude and longitude coordinates of the sampling center point.

After removing impurities such as plant residues and stones from the soil samples, they were placed in a thermostatic, electric-heating hot-air-drying oven for drying. Subsequently, the soil samples were ground using an agate mortar and sieved through a 200-mesh soil sieve. In the laboratory, we utilized the flame atomic absorption spectrophotometry method specified in Standard of the Ministry of Ecology and Environment of the People's Republic of China to determine the content of Cu in the soil samples [29]. Additionally, we employed the potentiometry method stipulated in another standard of the same department to measure the pH value of the soil [30].

2.3. Multispectral Data Acquisition and Pre-Processing

Although most existing studies primarily use images from the bare soil period as the ideal window for effectively predicting soil properties at a regional scale [31–33], obtaining images during the optimal bare soil window can be challenging, particularly in regions like southern China, where the bare soil period is short and the environment is cloudy and rainy [34]. Given these limitations, an increasing number of researchers have begun directly using remote sensing images to generate vegetation variables for soil mapping [35]. The close relationship between pollutant concentrations and plant growth variables, along with their spectral responses, allows for the detection and quantification of plant stress induced by heavy metals [13,36]. Therefore, this enhances the effectiveness of using satellite images for digital soil mapping in areas with minimal environmental gradients.

This study utilized Sentinel-2 Level-2A remote sensing imagery, downloaded through the Google Earth Engine (GEE) cloud platform. By defining a cloud and cirrus masking function within the Sentinel-2 images and using the QA60 band, which contains image quality assessment information, we selected an image collection between 1 April 2022 and 30 July 2022, with cloud cover below 5%. After filtering and masking, the image collection was averaged. This period corresponds to the growth stage of early rice from sowing to harvesting in the Zhuhai and Zhongshan regions. Although there may be some phenological differences in vegetation across different sampling points, the rice crops were generally at similar growth stages. Table 1 shows the spatial resolution and spectral range characteristics of the Sentinel-2 satellite's Multispectral Instrument (MSI). In this study, we used Sen2Cor-processed Level-2A products, which underwent atmospheric and orthorectification corrections to obtain surface reflectance data. To meet the research requirements, we used cubic convolution interpolation on the GEE platform to resample the Level-2A data's B1 and B5 to B12 bands at a 10 m resolution. This process helped ensure data consistency and accuracy, providing a more precise basis for exploring soil heavy metal characteristics in subsequent analyses.

Table 1. Sentinel-2 Level-2A MSI sensor parameters.

Band ID	Description	Spatial Resolution (m)	Wavelength (μm)
B1	Coastal Aerosol	60	0.433–0.453
B2	Blue	10	0.457–0.522
B3	Green	10	0.542–0.577
B4	Red	10	0.650–0.680
B5	Red edge 1	20	0.697–0.712
B6	Red edge 2	20	0.732–0.747
B7	Red edge 3	20	0.733–0.793
B8	NIR 1	10	0.784–0.899
B8A	NIR 2	20	0.855–0.875
B9	Water vapor	60	0.935–0.955
B11	SWIR 1	20	1.565–1.655
B12	SWIR 2	20	2.100–2.280

NIR: near-infrared; SWIR: shortwave infrared.

2.4. Construction of Spectral Indices in Two Dimensions (2D) and Three Dimensions (3D)

To mitigate the limitations of fewer multispectral bands and larger spectral intervals, spectral indices are constructed to extract more spectral information from multispectral data. According to previous studies, using spectral indices to invert farmland soil heavy metal content can yield better results [37]. Among these, 2D spectral indices utilize two spectral dimensions, focusing on the relationship between two specific bands, while three-dimensional (3D) spectral indices incorporate information from three spectral bands or wavelengths, providing a more complex and potentially more informative representation of spectral characteristics.

The selection of spectral indices is crucial for accurately reflecting soil heavy metal content, given that each spectral index is based on specific local environmental conditions [38]. Therefore, in this study, commonly used two-dimensional spectral index forms were selected, including difference indices (DIs), ratio indices (RIs), and normalized difference indices (NDIs). Reference was also made to the common three-dimensional spectral index forms proposed by Wang [17] and Cao [39] in the literature. By iterating through all possible spectral indices, the aim was to find the most suitable spectral index form for the study area. Table 2 details all the spectral index forms iterated in this study and their calculation formulas.

Table 2. Spectral index formulas in 2D and 3D.

Spectral indices	Description	Formula	
Spectral indices in two dimensions (2D)	Difference Indices	$DI(R_i, R_j) = R_i - R_j$	(1)
	Ratio Indices	$RI(R_i, R_j) = \frac{R_i}{R_j}$	(2)
	Normalized Differential Indices	$NDI(R_i, R_j) = \frac{R_i - R_j}{R_i + R_j}$	(3)
Spectral indices in three dimensions (3D)	TBI1	$TBI1 = \frac{R_i - R_j}{R_j + R_k}$	(4)
	TBI2	$TBI2 = \frac{R_i - R_j}{R_k}$	(5)
	TBI3	$TBI3 = (R_i - R_j) - (R_j - R_k)$	(6)
	TBI4	$TBI4 = \frac{R_i}{R_j + R_k}$	(7)
	TBI5	$TBI5 = \frac{R_i - R_j}{[(R_i - R_j) - (R_j - R_k)]}$	(8)

TBI_Z represents the spectral Indices of the Z-th transformation form, where $Z = 1, 2, \dots, 5$. R_i , R_j , and R_k are the reflectances of any three bands selected from all bands of the Sentinel-2 satellite.

2.5. Optimal Inversion Ranges for pH and Soil Sample Classification

Previous studies have shown that soil pH indirectly affects spectral reflectance, thereby affecting the accuracy of estimating heavy metal concentrations in the soil. In practical situations, there is no direct relationship between soil pH and heavy metal concentrations in the soil. Therefore, it is difficult to use it as a covariate to construct an inversion model

for Cu content in soil. In this regard, this study achieved the classification of soil samples by constructing a decision tree with the pH value as the branching criterion and determining the optimal pH range for the inversion of soil Cu content, thereby improving the accuracy of the inversion model.

In this study, soil samples were divided into different intervals based on pH values. All soil samples were divided into three datasets, with samples classified into the pH intervals of [pHmin, pH1), [pH1, pH2), and [pH2, pHmax]. Here, pHmin represents the minimum pH value observed in the soil samples, while pHmax represents the maximum pH value observed. Based on the above division, a decision tree was constructed, and the samples were assigned to different branches according to the division of each node. This process was repeated step by step until the number of samples in each branch was not less than 70. The threshold combinations for the pH value decision tree are shown in Table 3, according to the actual division of soil samples.

Table 3. Threshold combination of pH decision tree.

pH1	pH2	pH1	pH2
5.85	7.75~8.05	6.35	7.90~8.05
5.95	7.80~8.05	6.45	7.90~8.05
6.05	7.85~8.05	6.55	7.95~8.05
6.15	7.90~8.05	6.65	8.00~8.05
6.25	7.90~8.05	6.75	8.05

When selecting the optimal pH value decision tree for the inversion of Cu content, this study used the correlation coefficient of the pH decision tree (R_{ph}) as a key evaluation index to accurately measure the effectiveness of decision tree construction [40]. When this evaluation index reached its maximum value, the corresponding pH value decision tree was recognized as the best decision tree. The specific calculation method for R_{ph} is detailed in Equation (9).

$$R_{ph} = \sum_{n=1}^3 \max |R_n(B_i, S_{ph})| \quad (9)$$

In this equation, B_i represents the spectral reflectance of the i -th band of the Sentinel-2 satellite; S_{ph} represents the soil Cu content (mg/kg) at the specified pH value; and R_n represents the correlation coefficient between the spectral reflectance of the i -th band of the Sentinel-2 satellite within the n th branch and the soil Cu content at the specified pH value.

2.6. Spectral Feature Selection

Utilizing the raw bands of Sentinel-2, a substantial amount of two-dimensional and three-dimensional spectral index data have been generated, characterized by considerable redundancy and complex combinations. In the model construction process, redundant and interfering variables can impact the accuracy and precision of Cu content inversion, thereby necessitating variable selection. In this investigation, the Successive Projections Algorithm (SPA) combined with the Boruta algorithm was employed for feature selection.

2.6.1. Successive Projections Algorithm (SPA)

The SPA is an algorithm used for feature selection. It achieves dimensionality reduction by gradually selecting the most relevant spectral bands to construct a subset [41]. The specific algorithm steps are as follows: The spectral reflectance matrix is X , with m columns, and the initial iteration vector is $X_{k(0)}$. The number of selected spectral bands is N .

1. A column in matrix X is arbitrarily selected, denoted as the i -th column, and assigned to X_i , denoted as $X_{k(0)}$.
2. The set of positions of the remaining column vectors is denoted as S , where i represents the indices of all potential features, and k is the indices of the already selected features.

$$S = \{i, 1 \leq i \leq m, i \notin (k(0), \dots, k(N-1))\} \quad (10)$$

3. The projection of the currently selected variable $X_{k(N-1)}$ onto the column vectors X_i of the remaining original spectral data is calculated:

$$P_{X_i} = X_i - \left(X_i^T X_{k(N-1)} \right) X_{k(N-1)} \left(X_{k(N-1)}^T X_{k(N-1)} \right)^{-1}, i \in S \quad (11)$$

4. The maximum projection is determined.

$$k(n) = \arg(\max(\|P_{X_i}\|)), i \in S \quad (12)$$

5. The maximum projection vector is used as the projection vector for the next iteration.

$$X_i = P_{X_i}, i \in S \quad (13)$$

6. For $n = n + 1$, if $n < N$, then step (2) is repeated. Finally, the extracted variables are $\{X_{k(n)}; n = 0, \dots, N-1\}$

2.6.2. Boruta

The Boruta algorithm is a feature selection method based on Random Forest (RF). Its core principle involves identifying truly important features from a given set and eliminating redundant ones [42,43]. The specific algorithm steps are as follows:

1. Perform multiple bootstrap resampling iterations on the original dataset, where the original data are randomly replaced by a series of shadow features, creating a Random Forest model for each resampled dataset.
2. The Boruta algorithm introduces shadow features, which are generated by shuffling the order of the original features, adding noise, and randomizing them. These shadow features are used to simulate randomly selected features and are compared with the original features.
3. The importance scores of each original feature and its corresponding shadow features are computed and compared. A feature is considered important if its importance score is significantly higher than that of its shadow feature; otherwise, it is marked as unimportant.
4. The Boruta algorithm iteratively repeats the steps of calculating feature importance and comparison until all features are definitively classified as important or unimportant.
5. After several iterations, the original features that consistently show significantly higher importance than their corresponding shadow features are selected as important. These features are retained, while those deemed unimportant are removed.

2.7. Ensemble Learning

The core idea of ensemble learning is to improve overall prediction performance by combining the predictions of multiple base learners. Numerous previous studies have shown that ensemble models typically have higher accuracy than single models [44]. Among them, bagging and boosting are two common ensemble strategies. Bagging constructs multiple base regression models through bootstrap sampling and averages or votes on their predictions to reduce model variance and improve prediction stability. This method helps to reduce the risk of overfitting that a single model may produce. Boosting is a method that reduces bias in supervised learning. It trains a series of weak regression models and combines them into a strong regression model based on the prediction errors of each model. Boosting gradually focuses on the samples that the previous models failed to predict correctly, thus continuously improving prediction performance.

This study employed ensemble learning algorithms to construct a remote sensing inversion model for Cu content and selects the optimal prediction model. Two typical

decision-tree-based ensemble learning algorithms were chosen, including Random Forest (RF) using the bagging strategy and Extreme Gradient Boosting (XGBoost) using the boosting strategy.

2.7.1. Random Forest (RF)

RF is an ensemble learning method based on decision trees, which improves prediction accuracy and stability by constructing multiple decision trees and summarizing their prediction results. The core idea of RF [45] is to use the bagging ensemble strategy, which constructs multiple different decision trees by bootstrapping the training data and randomly selects a subset of features to consider during the node splitting process of each decision tree, thereby increasing the diversity and generalization ability of the model.

2.7.2. Extreme Gradient Boosting (XGBoost)

XGBoost is a machine learning algorithm based on the gradient boosting decision tree (GBDT), characterized by its efficiency, flexibility, and portability. Based on the boosting integration strategy [46], XGBoost introduces a series of optimization measures to improve the performance and efficiency of the model. The core idea of XGBoost is to construct a series of weak learners and gradually correct previous prediction errors to approximate the true objective function. In each iteration, XGBoost calculates the residuals of the current model and trains a new decision tree based on these residuals. The new decision tree focuses on optimizing the samples that the previous model failed to predict correctly, thus gradually reducing prediction errors.

2.8. Optuna-Based Parameter Tuning Framework

Optuna (v3.5) [47] is an automatic hyperparameter optimization software framework, particularly designed for machine learning. It automates the search for optimal hyperparameters, making it easier to fine-tune models. Optuna can accurately identify which hyperparameters have the most significant impact on the objective function, so it can serve as a guide for focusing on which hyperparameters should be analyzed further to achieve further optimization. Compared with a grid search, this algorithm adopts sampling and pruning strategies to select optimal hyperparameters, achieving an efficient search. Although the runtime of a grid search may be shorter than that of Optuna, manually constructing hundreds of grid search experiments is time-consuming. Meanwhile, Optuna automatically selects the next set of hyperparameters for testing, eliminating the need for manual intervention. More details about the aforementioned parameters as well as other parameters in the “Optuna” package are described in the official documentation for Optuna (<https://optuna.readthedocs.io/en/latest/index.html>, accessed on 15 June 2024).

This study combined the Optuna framework with the XGBoost and RF algorithms, aiming to invert the concentration of Cu content in the soil through the following steps: (1) An objective function was defined to evaluate the performance of the model and return the corresponding evaluation metrics. In this function, we explicitly defined the search space for hyperparameters. (2) When conducting individual experiments, we trained a regression model (including XGBoost and RF) using a calibration dataset, and then we made predictions using a validation dataset and calculated the root mean square error (RMSE) to assess prediction accuracy. (3) By conducting multiple experiments, the optimal parameter configuration of the model was determined based on the value of the loss function. In this study, the XGBoost algorithm required the adjustment of nine key parameters, as shown in Table 4. Meanwhile, the RF algorithm adjusted only one parameter, which was the number of trees (*n_estimators*). Due to the relatively small number of samples and features used in this study, the other parameters were set to default values.

Table 4. The range of values for XGBoost hyperparameters.

Parameter	Recommended Range	Data Type
num_boost_round	[500, 2000]	int
eta	[0.01, 0.3]	float
max_depth	[3, 10]	int
subsample	[0.5, 1]	float
colsample_bytree	[0.5, 1]	float
gamma	[0, 1]	float
min_child_weight	[1, 10]	float
rel_lambda	[0, 1]	float
alpha	[0, 1]	float

2.9. Evaluation Indicators

Three accuracy indicators were utilized to evaluate the accuracy of the regression models for soil Cu content: R^2 , RMSE, and RPD. Generally, a proficient model demonstrates high values for R^2 and RPD and low scores for RMSE [48]. Each indicator is computed as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} \quad (14)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (15)$$

$$RPD = \frac{SD}{RMSE} = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y}_i)^2}{\sum_{i=1}^n (y_i - \hat{y}_i)^2}} \quad (16)$$

In these equations, n represents the number of soil samples; y_i and \hat{y}_i denote the measured value and estimated heavy metal concentration of the i -th soil sample in the validation dataset, respectively; and \bar{y}_i represents the average heavy metal concentration.

3. Results

3.1. Descriptive Statistics of the Heavy Metal Content

The basic statistical characteristics of Cu content in 242 collected soil samples are shown in Table 5. The minimum Cu content in the soil is 3 mg/kg, the maximum is 271 mg/kg, the standard deviation is 29.7 mg/kg, and the coefficient of variation is 87.6%. The high coefficient of variation indicates a large degree of dispersion in soil Cu content in the samples. Meanwhile, the average Cu content in the soil in the study area is 33.9 mg/kg, significantly higher than the background values of surface soil in Guangdong Province and China, and three times that of the surface soil background value in Guangdong Province [49].

Table 5. Statistical description of Cu concentration for soil samples collected.

Metal	Number	Minimum	Maximum	Mean	Standard Deviation	Coefficient of Variation (%)	Guangdong Soil Background Value	Chinese Soil Background Value
Cu	242	3	271	33.9	29.7	87.6	11.2	20

3.2. Analysis of the Impact of Environmental Variables on the Spectrum

The overall samples of the study area were divided into four intervals based on the pH value, and the mean spectral reflectance of each type of soil was calculated to obtain the average spectral reflectance of the soil under different pH conditions, as shown in Figure 3. Although the spectral reflectance curves of soil under different pH conditions are different, the transformation trends are relatively similar. A distinct absorption valley can be observed in the red band range of the visible light spectrum. The spectral reflectance

gradually increases with wavelength in the range from B1 to B8A, but gradually decreases after B8A. It is noteworthy that the reflectance significantly increases with the rise in soil pH in the B1 to B6 and B11 to B12 bands. However, there is a clear negative correlation between reflectance and soil pH in the B6 to B9 bands. Soils with different alkalinity levels exhibit high distinguishability in the spectrum, demonstrating that alkaline soils have unique and distinct spectral response characteristics in each band of the Sentinel-2 satellite.

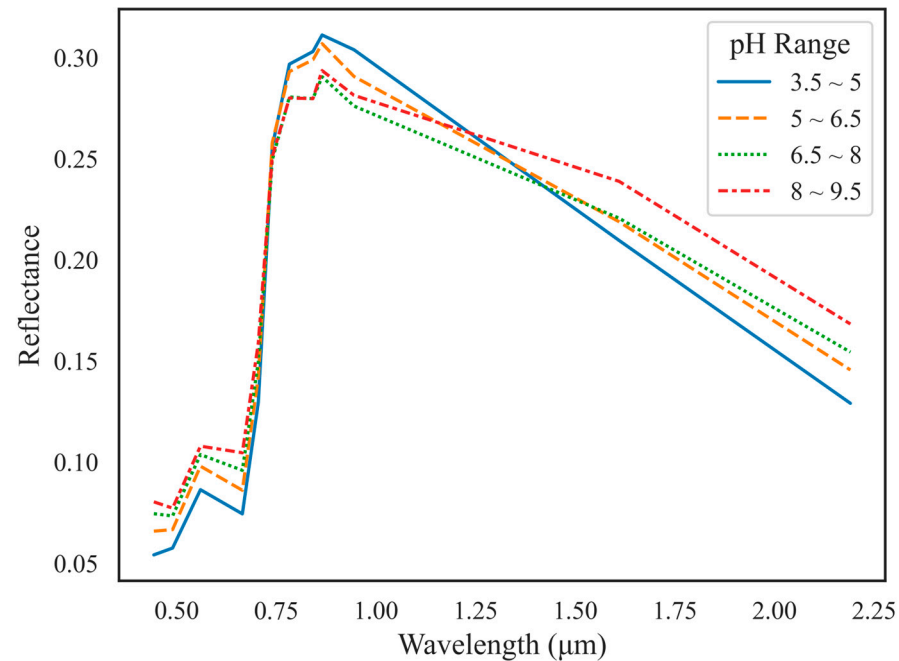


Figure 3. Average spectral reflectance of soil under different pH conditions.

3.3. Determining the Optimal Decision Tree

By constructing a soil pH decision tree to determine the category of each soil sample, the sensitivity of spectral reflectance to soil Cu content is enhanced. The construction results, shown in Figure 4, indicate that when pH1 and pH2 are 5.85 and 7.75, respectively, the Rph values of all bands reach their maximum, identifying the optimal pH decision tree. Furthermore, according to the classification results, the pH ranges for the soil samples in the three different categories are [3.63, 5.85), [5.85, 7.75), and [7.75, 9.25], respectively.

3.4. Spectral Feature Extraction

The experiment utilized Sentinel-2 original bands and their varying spectral indices in two and three dimensions, obtaining 7008 feature variables. During the variable selection process, we initially used the SPA to select the top 100 most important bands, thereby reducing the number of features and lowering collinearity among them. Subsequently, we applied the Boruta algorithm to further refine these 100 bands. The Boruta algorithm generates a set of shadow features by shuffling the values of the original features, rendering them irrelevant. These original and shadow features are then combined to train the RF model, allowing for the calculation of importance for each feature. This method not only comprehensively considers all features that significantly impact the prediction target but also effectively identifies important features in the variable selection process, rather than merely finding an optimal subset.

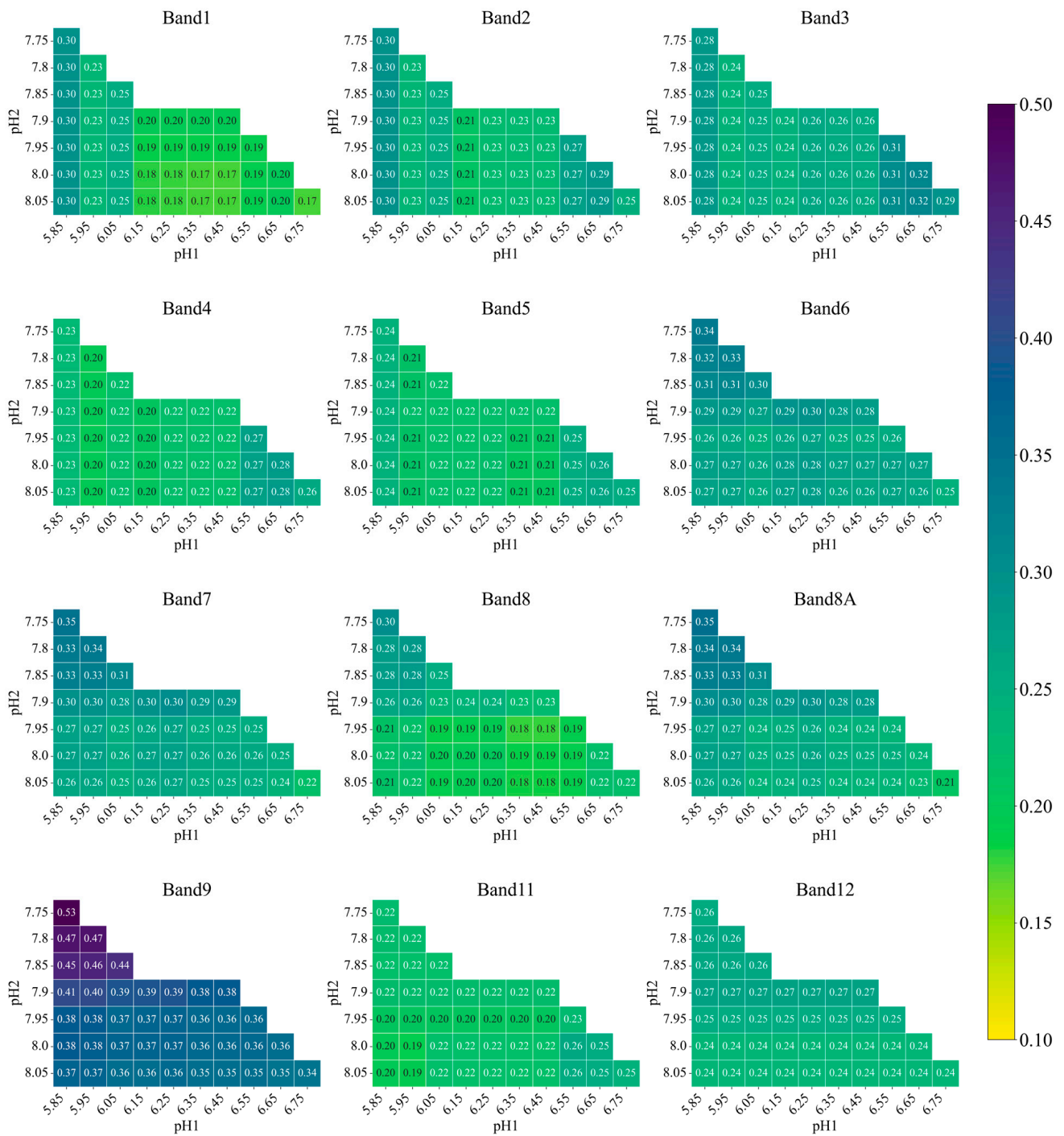


Figure 4. Construction of Cu content decision tree with pH as branching criterion.

Table 6 shows the results of the feature band selection conducted both with and without considering the decision tree. It is noteworthy that, in all cases, no variations in the two-dimensional spectral indices are selected. This may be because the predictive ability of the two-dimensional spectral indices is inferior to that of the three-dimensional spectral indices. Additionally, the selected feature bands mainly focus on the TB12 and TB14 of the three-dimensional spectral indices, indicating that variations in these indices have a high correlation with and provide more information for the inversion of soil heavy metal Cu.

Table 6. Comparison of feature band selection results under different branching conditions.

Branch Criteria	pH			
	/	[3.63, 5.85)	[5.85, 7.75)	[7.75, 9.25]
Content range	Full	[3.63, 5.85)	[5.85, 7.75)	[7.75, 9.25]
Feature Bands	B1	B1	TBI2_B2_B8_B7	TBI2_B8_B2_B1
	TBI4_B8_B7_B6	TBI4_B2_B12_B10	TBI2_B4_B8_B7	TBI4_B9_B11_B2
	TBI2_B9_B2_B1	TBI2_B9_B3_B4	TBI2_B12_B7_B8	TBI4_B1_B11_B8
	TBI2_B2_B10_B11	TBI4_B5_B11_B9	TBI4_B4_B6_B10	TBI4_B12_B6_B11
	TBI2_B11_B8_B7	TBI4_B4_B12_B6	TBI4_B7_B11_B2	TBI4_B6_B7_B8
	TBI4_B9_B11_B2	TBI4_B10_B12_B2	TBI4_B4_B12_B6	TBI2_B10_B1_B4
	TBI4_B4_B12_B6	TBI4_B6_B12_B2	TBI4_B10_B6_B11	TBI2_B10_B2_B1
	TBI4_B4_B6_B10		TBI4_B5_B11_B8	TBI2_B10_B1_B2
	TBI2_B3_B8_B7		TBI4_B11_B12_B4	
	TBI4_B12_B11_B8		TBI2_B10_B7_B8	
	TBI4_B2_B11_B9		TBI2_B10_B8_B7	

3.5. Importance Analysis of Hyperparameters

In this study, various scenarios were experimented with using the Optuna automatic hyperparameter optimization algorithm, and the corresponding optimization results were recorded, as shown in Table 7. The Optuna algorithm evaluates the influence of each hyperparameter on the RMSE loss function. In Figure 5, the influence of each hyperparameter in the XGBoost model on the RMSE loss function is demonstrated without constructing a decision tree and over different pH ranges. The results show that, under any condition, the three hyperparameters “min_child_weight”, “eta”, and “subsample” are always the most important. Therefore, adjusting these three hyperparameters is crucial for improving the predictive performance of Cu content in soil.

Table 7. Optimization results of hyperparameters under different conditions.

Branch Criteria	Content Range	RF			XGBoost						
		A	B	C	D	E	F	G	H	I	J
/	Full	74	0.015683	9	0.715672	0.680002	5.450774	0.017943	1900	0.165201	0.119072
	pH < 5.85	290	0.149107	6	0.953130	0.884508	7.275414	0.010593	988	0.416995	0.268549
pH	5.85 ≤ pH < 7.75	805	0.978662	8	0.568985	0.984669	9.918633	0.260176	1585	0.190610	0.131338
	pH ≥ 7.75	983	0.042852	3	0.813619	0.742660	7.182743	0.231286	1376	0.054283	0.796611

A: n_estimators; B: gamma; C: max_depth; D: subsample; E: colsample_bytree; F: min_child_weight; G: eta; H: num_boost_round; I: alpha; J: rel_lambda.

In the XGBoost model, adjusting the “min_child_weight” parameter is a key factor in balancing model complexity and generalization ability. This parameter specifies the minimum sum of sample weights in a leaf node, controlling the growth of the tree. The “eta” parameter adjusts the learning rate for each tree during each iteration, thereby influencing the convergence speed and robustness of the model. The “subsample” parameter is used to control the sampling ratio of training samples for each tree. By setting the “subsample” parameter, the model can randomly select a portion of the training data in each training round, which helps the model to avoid relying too heavily on specific samples, thereby improving the robustness and performance of the model. Through the Optuna automatic hyperparameter optimization algorithm, we can gain insight into which hyperparameters have a more significant impact on the model’s performance, and thus achieve further optimization.

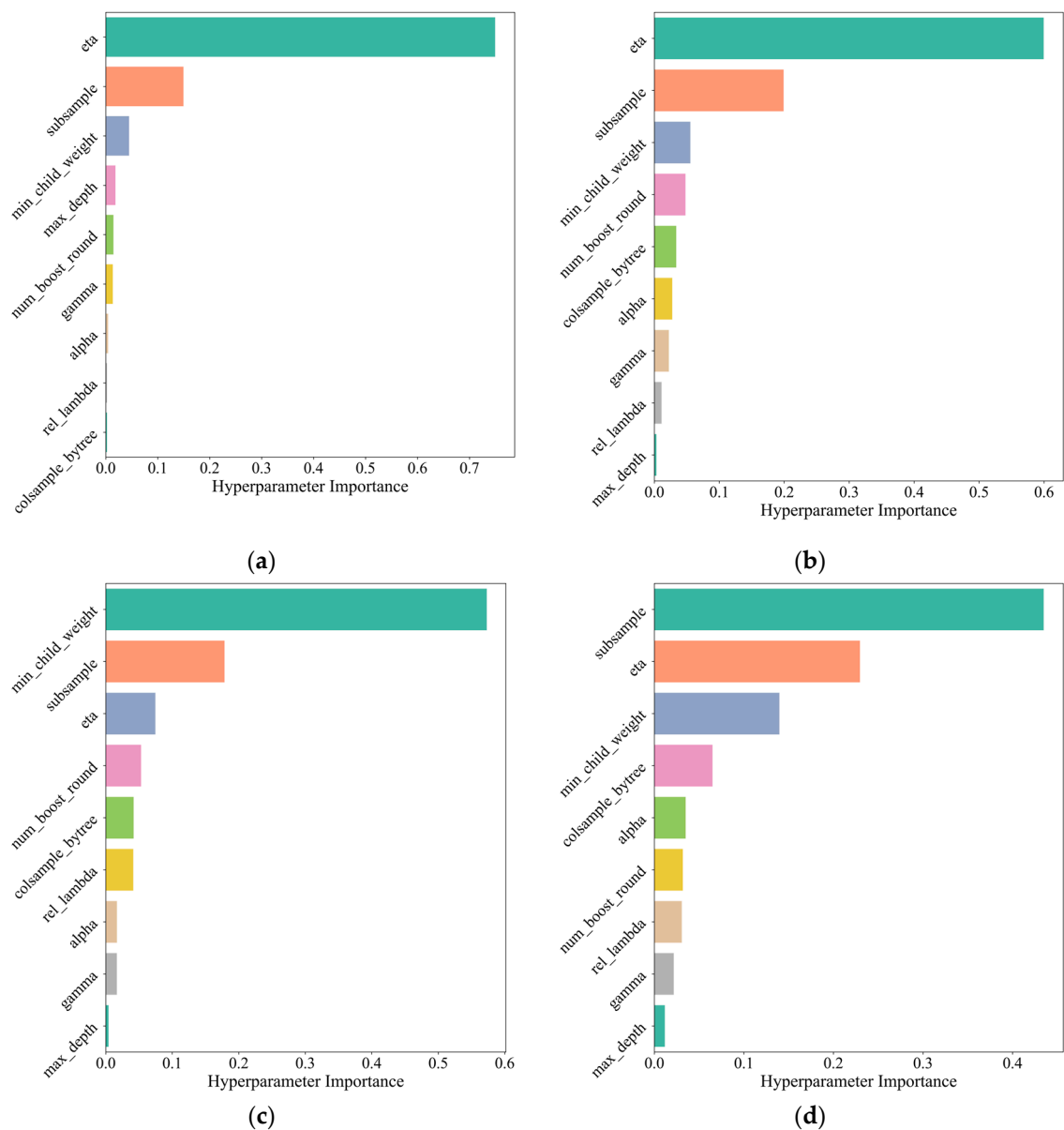


Figure 5. Assessment of the importance of hyperparameters. (a) Full; (b) $\text{pH} < 5.85$; (c) $5.85 \leq \text{pH} < 7.75$; and (d) $\text{pH} \geq 7.75$.

3.6. Comparison of Inversion Accuracy

In order to compare the accuracy of the RF model based on the bagging strategy and the XGBoost model based on the boosting strategy in the inversion of heavy metal Cu content, the performance of the models was comprehensively evaluated using the R^2 , RMSE, and RPD indicators. Table 8 presents the evaluation results of soil Cu content inversion without constructing decision trees and under different soil pH conditions. When a decision tree was not constructed, the inversion of Cu content in the soil basically lacked predictive ability. When considering a decision tree, as shown in Figure 6, when the soil pH value is between 5.85 and 7.75, the R^2 of the RF model in the validation set is 0.54, and that of the XGBoost model is 0.76, the corresponding RMSEs are 22.48 and 16.12, and the RPDs are 1.51 and 2.11, respectively. In contrast, the XGBoost model performed better under the same conditions. The R^2 and RPD of the XGBoost model increased by 0.22 and 0.6, respectively, while the RMSE value decreased by 6.36. This result indicates that when the soil pH value is between 5.85 and 7.75, the XGBoost model can exhibit higher prediction accuracy compared to the RF model. The XGBoost model exhibited excellent performance

in this range, indicating that decision trees can effectively enhance the sensitivity of spectral indices to soil Cu content, thereby significantly improving inversion accuracy. Additionally, when the pH is below 5.85, the R^2 values for the RF and XGBoost models are -0.44 and 0.2, respectively. When the pH is above 7.75, the R^2 values are 0.43 and 0.5, respectively. This indicates that within these pH ranges, the relationship between spectral reflectance and soil Cu content is weak or even misleading, which in turn affects the accuracy of predicting soil Cu content.

Table 8. Evaluation of inversion results of soil Cu content at different soil pH values.

Regression Methods	Soil pH	Calibration Dataset			Validation Dataset		
		R^2	RMSE	RPD	R^2	RMSE	RPD
RF	Full	0.63	20.17	1.65	0.17	15.65	1.10
	pH < 5.85	0.89	7.38	3.03	-0.44	18.45	0.85
	5.85 ≤ pH < 7.75	0.87	12.95	2.75	0.54	22.48	1.51
	pH ≥ 7.75	0.61	21.36	1.61	0.43	13.84	1.35
XGBoost	Full	0.40	25.72	1.30	0.33	14.04	1.23
	pH < 5.85	0.17	20.18	1.11	0.20	13.71	1.15
	5.85 ≤ pH < 7.75	0.70	19.38	1.84	0.76	16.12	2.11
	pH ≥ 7.75	0.58	22.05	1.56	0.50	12.92	1.44

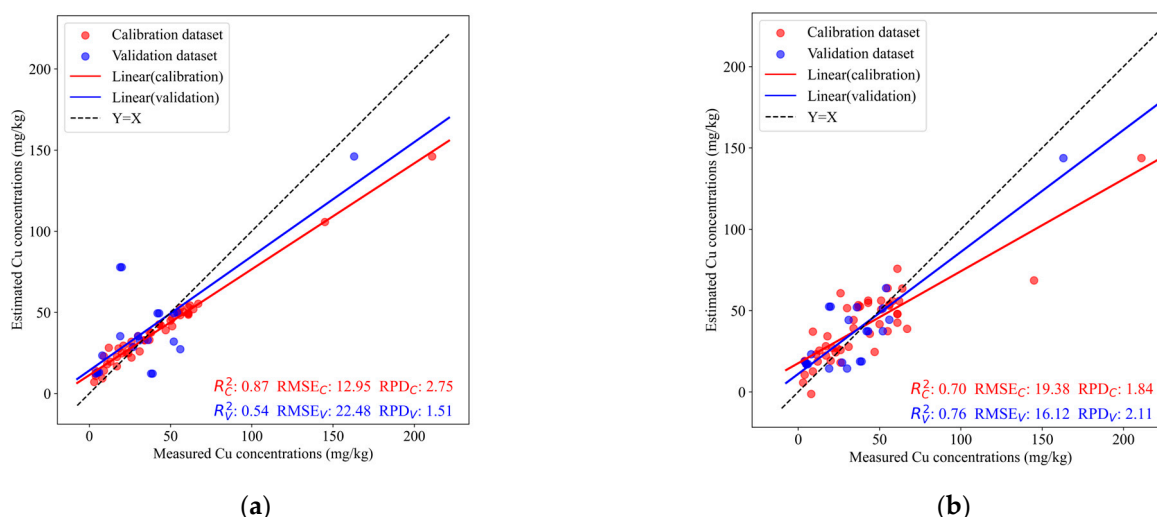


Figure 6. Scatter plot of the measured and estimated values with pH values between 5.85 and 7.75: (a) RF; (b) XGBoost.

Overall, the accurate evaluation of decision trees constructed based on pH as a branching criterion revealed that the feasibility and reliability of using Sentinel-2 satellite multi-spectral images to retrieve soil Cu concentrations are acceptable.

To validate the effectiveness of the model, white noise data were randomly generated and combined with actual Cu content to train and evaluate the model. Specifically, white noise data with the same dimensions as the original data (7008 features, 242 samples) were generated as independent variables. Feature variables were selected using a combined method of the SPA and the Boruta algorithm. The optimal RMSE loss function was determined through Optuna automatic hyperparameter optimization, and the same RF and XGBoost regression models were used for training. As shown in Table 9, the performance of the model on white noise data is significantly lower than its performance on real data, with all R^2 values being negative. This indicates that the model can identify and utilize valid information in real data, rather than relying on random noise. The features selected by the SPA and the Boruta algorithm did not improve the model performance on white noise data, further validating the importance and effectiveness of these features in real data.

Table 9. Evaluation of Soil Cu Content Inversion Results on White Noise Data and Real Data.

Data Type	Soil pH	Regression Methods	Calibration Dataset			Validation Dataset		
			R ²	RMSE	RPD	R ²	RMSE	RPD
Real data	Optimal branch	RF	0.87	12.95	2.75	0.54	22.48	1.51
		XGBoost	0.70	19.38	1.84	0.76	16.12	2.11
White noise data	/	RF	0.84	13.54	2.47	−0.25	19.39	0.90
		XGBoost	0.40	25.88	1.29	−0.15	18.61	0.94

4. Discussion

4.1. Exploring the Potential of Sentinel-2 MSI in Soil Heavy Metal Retrieval

The limitations of multispectral data may result in the inversion accuracy of heavy metals based on Sentinel-2 being lower than that of hyperspectral data. However, by combining two-dimensional and three-dimensional spectral indices, it is possible to partially compensate for the limitations of the multispectral data's band range. This enhances the accuracy of heavy metal inversion by integrating multiple bands and effectively addressing the nonlinear relationships between spectral characteristics and soil heavy metal content. In this study, a decision tree with pH values as the branching criterion was constructed, significantly improving the inversion capacity for Cu content in the soil. Without the decision tree model, the predictive ability for Cu content in soil was weak; however, after introducing the decision tree, the inversion accuracy was greatly enhanced when pH values ranged between 5.85 and 7.75. Outside of this range, the relationship between spectral reflectance and soil Cu content became weak or even misleading, further affecting the accuracy of Cu content prediction. Additionally, while many studies incorporate geographical factors when using multispectral data for soil heavy metal inversion [15,17], this study achieved good inversion results solely by utilizing multispectral data and constructing a decision tree. Therefore, the use of Sentinel-2 and even other multispectral data holds great potential for estimating heavy metal content in regional soil surfaces.

4.2. Research Limitations and Future Work

This study determined the optimal pH range for the inversion of soil Cu content by constructing a decision tree based on pH as the branching criterion, thereby enhancing the sensitivity of spectral indices to the Cu content of soil within this range. However, only 242 sampling points were collected in this study, so the obtained optimal pH range may lack representativeness. In future research, it is recommended to increase the number of sampling points to improve the universality of the experimental results. Furthermore, future research could not only classify based on pH values but also explore constructing decision trees using moisture content or soil depth as classification criteria to further optimize the inversion model for estimating heavy metal content. To address inconsistencies between moisture content in remote sensing images and ground truth measurements, the tasseled cap transformation can be used to extract the wetness component, which can then serve as a branching criterion. The relationship between soil Cu content and spectral reflectance varies significantly across different soil depths, and this variation is associated with the Normalized Difference Vegetation Index (NDVI). By constructing decision trees with the NDVI as the branching criterion, the optimal depth for Cu content inversion can be determined. Additionally, future research could consider multispectral data fusion methods, such as combining data from Landsat 8 OLI and Sentinel-2 MSI for soil heavy metal inversion studies. This approach is expected to improve the accuracy and reliability of inversion models.

5. Conclusions

In response to the issue of low accuracy in the multispectral inversion of soil heavy metals, this study employed Sentinel-2 multispectral data to construct two-dimensional

and three-dimensional spectral indices. Using soil pH values as a branching criterion, decision trees were built to classify sample categories, enhancing the sensitivity of spectral indices to soil heavy metal Cu content. Based on this, this study utilized the SPA combined with the Boruta algorithm to select the characteristic bands of each branch and combined it with Optuna automatic hyperparameter optimization to use ensemble learning models to invert the Cu content in the soil. The main conclusions of this study are as follows:

1. Constructing a soil pH decision tree significantly enhances the sensitivity of spectral indices to soil heavy metal Cu, thereby effectively improving the inversion accuracy. According to the results of the inversion accuracy study, soil over-acidity and over-alkalinity significantly impact the experimental results. In this study, when the pH is between 5.85 and 7.75, the influence of soil pH on spectral indices is minimal, thereby achieving the highest inversion accuracy.
2. Based on Optuna combined with ensemble learning models, it was found that under any conditions for predicting the concentration of Cu in soil, the three hyperparameters 'min_child_weight', 'eta', and 'subsample' have a significant impact on the RMSE loss function of the XGBoost model. Therefore, optimizing these parameters should be a priority when using multispectral data to predict Cu concentration in soil.
3. By constructing two-dimensional and three-dimensional spectral indices and applying the SPA combined with the Boruta algorithm for feature spectral index selection, it was found that the characteristic spectral indices used to invert soil heavy metal Cu content are mainly concentrated in the spectral transformation forms of TBI2 and TBI4. Additionally, through the construction of white noise, it was found that the model performance on white noise data is significantly lower than its performance on real data, further validating the importance and effectiveness of these characteristic spectral indices in real data.

Author Contributions: Conceptualization, F.W. and K.Y.; data curation, D.C.; formal analysis, K.Y.; funding acquisition, H.G.; methodology, F.W. and Z.C.; project administration, H.G.; software, Z.Y.; supervision, H.G.; validation, Z.Y., Z.C. and D.C.; visualization, F.W.; writing—original draft, F.W.; writing—review and editing, R.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation-Guangdong Joint Fund Key Project (Grant Number: U1911202), Key R&D Plan of the Ministry of Science and Technology (Grant Number: 2019YFC1805300), and Guangdong Provincial Natural Science Foundation Project (Grant Number: 2019A1515012131).

Data Availability Statement: The data that support the finding of this study are available from the corresponding author upon reasonable request. Due to privacy and other restrictions, these data are not publicly available.

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

References

1. Lin, H.; Wang, Z.; Liu, C.; Dong, Y. Technologies for Removing Heavy Metal from Contaminated Soils on Farmland: A Review. *Chemosphere* **2022**, *305*, 135457. [[CrossRef](#)] [[PubMed](#)]
2. Tsvetkov, P.; Coy, S.; Petrova, B.; Dreishpoon, M.; Verma, A.; Abdusamad, M.; Rossen, J.; Joesch-Cohen, L.; Humeidi, R.; Spangler, R.D.; et al. Copper Induces Cell Death by Targeting Lipoylated TCA Cycle Proteins. *Science* **2022**, *375*, 1254–1261. [[CrossRef](#)] [[PubMed](#)]
3. Lu, B.; Dao, P.D.; Liu, J.; He, Y.; Shang, J. Recent Advances of Hyperspectral Imaging Technology and Applications in Agriculture. *Remote Sens.* **2020**, *12*, 2659. [[CrossRef](#)]
4. Lou, P.; Fu, B.; He, H.; Chen, J.; Wu, T.; Lin, X.; Liu, L.; Fan, D.; Deng, T. An Effective Method for Canopy Chlorophyll Content Estimation of Marsh Vegetation Based on Multiscale Remote Sensing Data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 5311–5325. [[CrossRef](#)]
5. Pan, F.; Guo, J.; Miao, J.; Xu, H.; Tian, B.; Gong, D.; Zhao, J.; Lan, Y. Summer Maize LAI Retrieval Based on Multi-Source Remote Sensing Data. *Int. J. Agric. Biol. Eng.* **2023**, *16*, 179–186. [[CrossRef](#)]
6. Zhao, W.; Zhou, C.; Zhou, C.; Ma, H.; Wang, Z. Soil Salinity Inversion Model of Oasis in Arid Area Based on UAV Multispectral Remote Sensing. *Remote Sens.* **2022**, *14*, 1804. [[CrossRef](#)]

7. Zhao, W.; Ma, H.; Zhou, C.; Zhou, C.; Li, Z. Soil Salinity Inversion Model Based on BPNN Optimization Algorithm for UAV Multispectral Remote Sensing. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 6038–6047. [[CrossRef](#)]
8. Žižala, D.; Minařík, R.; Zádorová, T. Soil Organic Carbon Mapping Using Multispectral Remote Sensing Data: Prediction Ability of Data with Different Spatial and Spectral Resolutions. *Remote Sens.* **2019**, *11*, 2947. [[CrossRef](#)]
9. Wang, S.; Gao, J.; Zhuang, Q.; Lu, Y.; Gu, H.; Jin, X. Multispectral Remote Sensing Data Are Effective and Robust in Mapping Regional Forest Soil Organic Carbon Stocks in a Northeast Forest Region in China. *Remote Sens.* **2020**, *12*, 393. [[CrossRef](#)]
10. Xu, Y.; Smith, S.E.; Grunwald, S.; Abd-Elrahman, A.; Wani, S.P. Incorporation of Satellite Remote Sensing Pan-Sharpener Imagery into Digital Soil Prediction and Mapping Models to Characterize Soil Property Variability in Small Agricultural Fields. *ISPRS J. Photogramm. Remote Sens.* **2017**, *123*, 1–19. [[CrossRef](#)]
11. Li, X.; Cao, S.; Bai, X.; Li, H. Research progress of multispectral technology in soil composition content detection. *Spectrosc. Spectr. Anal.* **2020**, *40*, 2042–2047.
12. 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](#)]
13. Wang, F.; 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](#)]
14. Song, Y.; Sun, N.; Zhang, L.; Wang, L.; Su, H.; Chen, Z.; Yu, H.; Li, B. Using Multispectral Variables to Estimate Heavy Metals Content in Agricultural Soils: A Case of Suburban Area in Tianjin, China. *Geoderma Reg.* **2022**, *29*, e00540. [[CrossRef](#)]
15. Yu, H.; Xie, S.; Liu, P.; Hua, Z.; Song, C.; Jing, P. Estimation of Pb and Cd Content in Soil Using Sentinel-2A Multispectral Images Based on Ensemble Learning. *Remote Sens.* **2023**, *15*, 2299. [[CrossRef](#)]
16. Yang, N.; Han, L.; Liu, M. Inversion of Soil Heavy Metals in Metal Tailings Area Based on Different Spectral Transformation and Modeling Methods. *Heliyon* **2023**, *9*, e19782. [[CrossRef](#)]
17. Wang, Z.; Zhang, F.; Zhang, X.; Chan, N.W.; Kung, H.; Ariken, M.; Zhou, X.; Wang, Y. Regional Suitability Prediction of Soil Salinization Based on Remote-Sensing Derivatives and Optimal Spectral Index. *Sci. Total Environ.* **2021**, *775*, 145807. [[CrossRef](#)]
18. Nawar, S.; Mouazen, A.M. Predictive Performance of Mobile Vis-near Infrared Spectroscopy for Key Soil Properties at Different Geographical Scales by Using Spiking and Data Mining Techniques. *Catena* **2017**, *151*, 118–129. [[CrossRef](#)]
19. Reddy, K.J.; Wang, L.; Gloss, S.P. Solubility and Mobility of Copper, Zinc and Lead in Acidic Environments. *Plant Soil* **1995**, *171*, 53–58. [[CrossRef](#)]
20. Zhang, F.; Xiong, H.; Luan, F.; Lu, W. Measured spectral response characteristics of soil alkalization. *J. Infrared Millim. Terahertz Waves* **2011**, *30*, 55–60.
21. Zhang, C.; Gao, L.; Yun, W.; Li, L.; Ji, W.; Ma, J. Analysis of research progress on obtaining farmland quality evaluation indicators using remote sensing technology. *Trans. Chin. Soc. Agric. Mach.* **2022**, *53*, 1–13.
22. Jiang, Y.; Wang, J.; Yang, J. Analysis on the research progress of the cultivated land quality evaluation index system in black soil areas based on remote sensing technology. *Eng. Surv. Plan. Map Portal* **2023**, *32*, 1–7.
23. Ran, Y.; Ma, M.; Liu, Y.; Zhu, K.; Yi, X.; Wang, X.; Wu, S.; Huang, P. Physicochemical Determinants in Stabilizing Soil Aggregates along a Hydrological Stress Gradient on Reservoir Riparian Habitats: Implications to Soil Restoration. *Ecol. Eng.* **2020**, *143*, 105664. [[CrossRef](#)]
24. Kumar, S.; Bansod, B.S.; Thakur, R.; Jharwal, M.K. Soil pH Sensing Techniques and Technologies A Review. *Int. J. Adv. Res. Electr. Electron. Instrum. Eng.* **2015**, *4*, 4452–4456.
25. Li, S.; Li, C.; Cheng, S.; Xu, D.; Shi, Z. Soil profile organic carbon prediction based on field visible near-infrared spectroscopy and moisture effect correction algorithm. *Spectrosc Spect Anal* **2021**, *41*, 1234–1239.
26. Franceschini, M.H.D.; Dematté, J.A.M.; Kooistra, L.; Bartholomeus, H.; Rizzo, R.; Fongaro, C.T.; Molin, J.P. Effects of External Factors on Soil Reflectance Measured On-the-Go and Assessment of Potential Spectral Correction through Orthogonalisation and Standardisation Procedures. *Soil Tillage Res.* **2018**, *177*, 19–36. [[CrossRef](#)]
27. Peng, H.; Xia, H.; Chen, H.; Zhi, P.; Xu, Z. Spatial Variation Characteristics of Vegetation Phenology and Its Influencing Factors in the Subtropical Monsoon Climate Region of Southern China. *PLoS ONE* **2021**, *16*, e0250825. [[CrossRef](#)]
28. Xu, R.; Zhao, A.; Li, Q.; Kong, X.; Ji, G. Acidity Regime of the Red Soils in a Subtropical Region of Southern China under Field Conditions. *Geoderma* **2003**, *115*, 75–84. [[CrossRef](#)]
29. HJ 491-2019; Soil and Sediment—Determination of Copper, Zinc, Lead, Nickel and Chromium—Flame Atomic Absorption Spectrophotometry. Ministry of Ecology and Environment of the People’s Republic of China: Beijing, China, 2019.
30. HJ 962-2018; Soil—Determination of pH—Potentiometry. Ministry of Ecology and Environment of the People’s Republic of China: Beijing, China, 2018.
31. Geng, J.; Lv, J.; Pei, J.; Liao, C.; Tan, Q.; Wang, T.; Fang, H.; Wang, L. Prediction of Soil Organic Carbon in Black Soil Based on a Synergistic Scheme from Hyperspectral Data: Combining Fractional-Order Derivatives and Three-Dimensional Spectral Indices. *Comput. Electron. Agric.* **2024**, *220*, 108905. [[CrossRef](#)]
32. Guo, H.; Yang, K.; Wu, F.; Chen, Y.; Shen, J. Regional Inversion of Soil Heavy Metal Cr Content in Agricultural Land Using Zhuhai-1 Hyperspectral Images. *Sensors* **2023**, *23*, 8756. [[CrossRef](#)]
33. Liu, Q.; He, L.; Guo, L.; Wang, M.; Deng, D.; Lv, P.; Wang, R.; Jia, Z.; Hu, Z.; Wu, G.; et al. Digital Mapping of Soil Organic Carbon Density Using Newly Developed Bare Soil Spectral Indices and Deep Neural Network. *Catena* **2022**, *219*, 106603. [[CrossRef](#)]

34. Zepp, S.; Heiden, U.; Bachmann, M.; Wiesmeier, M.; Steininger, M.; van Wesemael, B. Estimation of Soil Organic Carbon Contents in Croplands of Bavaria from SCMaP Soil Reflectance Composites. *Remote Sens.* **2021**, *13*, 3141. [[CrossRef](#)]
35. Geng, J.; Tan, Q.; Lv, J.; Fang, H. Assessing Spatial Variations in Soil Organic Carbon and C:N Ratio in Northeast China's Black Soil Region: Insights from Landsat-9 Satellite and Crop Growth Information. *Soil Tillage Res.* **2024**, *235*, 105897. [[CrossRef](#)]
36. Wang, D.; Laffan, S.W.; Zhang, J.; Zhang, S.; Li, X. Quantitative Inversion of Soil Trace Elements from Spectroscopic Effects across Multiple Crop Growth Periods. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *132*, 104059. [[CrossRef](#)]
37. Chen, L.; Lai, J.; Tan, K.; Wang, X.; Chen, Y.; Ding, J. Development of a Soil Heavy Metal Estimation Method Based on a Spectral Index: Combining Fractional-Order Derivative Pretreatment and the Absorption Mechanism. *Sci. Total Environ.* **2022**, *813*, 151882. [[CrossRef](#)]
38. Viña, A.; Gitelson, A.A.; Nguy-Robertson, A.L.; Peng, Y. Comparison of Different Vegetation Indices for the Remote Assessment of Green Leaf Area Index of Crops. *Remote Sens. Environ.* **2011**, *115*, 3468–3478. [[CrossRef](#)]
39. Cao, Y.; Bao, N.; Zhou, B.; Gu, X.; Liu, S.; Yu, M. Research on Remote Sensing Inversion Method of Surface Water Content of Iron Tailings Based on Measured Spectra and Domestic Gaofen-5 Hyperspectral Satellite. *Spectrosc Spect Anal* **2023**, *43*, 1225–1233.
40. Du, R.; Cheng, J.; Zhang, Z.; Xu, Y.; Zhang, X.; Yin, H.; Yang, N. Sentinel-2 Multispectral Satellite Remote Sensing Retrieval of Soil Salinity Changes under Vegetation Cover. *Trans. Chin. Soc. Agric. Eng.* **2021**, *37*, 107–115.
41. Ye, S.; Wang, D.; Min, S. Successive Projections Algorithm Combined with Uninformative Variable Elimination for Spectral Variable Selection. *Chemom. Intell. Lab. Syst.* **2008**, *91*, 194–199. [[CrossRef](#)]
42. Zhou, J.; Zhang, R.; Guo, J.; Dai, J.; Zhang, J.; Zhang, L.; Miao, Y. Estimation of Aboveground Biomass of Senescence Grassland in China's Arid Region Using Multi-Source Data. *Sci. Total Environ.* **2024**, *918*, 170602. [[CrossRef](#)]
43. Nian, Y.; Su, X.; Yue, H.; Zhu, Y.; Li, J.; Wang, W.; Sheng, Y.; Ma, Q.; Liu, J.; Li, X. Estimation of the Rice Aboveground Biomass Based on the First Derivative Spectrum and Boruta Algorithm. *Front. Plant Sci.* **2024**, *15*, 1396183. [[CrossRef](#)] [[PubMed](#)]
44. Yang, Z.; He, Q.; Miao, S.; Wei, F.; Yu, M. Surface Soil Moisture Retrieval of China Using Multi-Source Data and Ensemble Learning. *Remote Sens.* **2023**, *15*, 2786. [[CrossRef](#)]
45. Rodriguez-Galiano, V.; Sanchez-Castillo, M.; Chica-Olmo, M.; Chica-Rivas, M. Machine Learning Predictive Models for Mineral Prospectivity: An Evaluation of Neural Networks, Random Forest, Regression Trees and Support Vector Machines. *Ore Geol. Rev.* **2015**, *71*, 804–818. [[CrossRef](#)]
46. Ye, M.; Zhu, L.; Li, X.; Ke, Y.; Huang, Y.; Chen, B.; Yu, H.; Li, H.; Feng, H. Estimation of the Soil Arsenic Concentration Using a Geographically Weighted XGBoost Model Based on Hyperspectral Data. *Sci. Total Environ.* **2023**, *858*, 159798. [[CrossRef](#)]
47. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-Generation Hyperparameter Optimization Framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 2623–2631.
48. Peng, J.; Biswas, A.; Jiang, Q.; Zhao, R.; Hu, J.; Hu, B.; Shi, Z. Estimating Soil Salinity from Remote Sensing and Terrain Data in Southern Xinjiang Province, China. *Geoderma* **2019**, *337*, 1309–1319. [[CrossRef](#)]
49. Chen, J.; Zhang, H.; Liu, J.; Li, F. Analysis on the spatial distribution characteristics of heavy metal elements in the soil surface under the regional geological background of Guangdong Province and its influencing factors. *Ecol. Environ. Sci.* **2011**, *20*, 646–651.

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.