

Article

Enhancing the Estimation and Mapping of Soil Cadmium by Using Geospatial Information-Guided Machine Learning and Principal Component Spectra

Jianfei Cao ^{1,2} , Yihui Liu ³ , Xinrong Duan ⁴, Xiaoli Liu ⁵, Xiukun Zhang ¹, Qixin Shi ¹ and Xibo Xu ^{1,6,*} ¹ College of Geography and Environment, Shandong Normal University, Jinan 250300, China; caojianfei@sdu.edu.cn (J.C.)² Satellite Application Center for Ecology and Environment, Ministry of Ecology and Environment, Beijing 100094, China³ Urban and Rural Planning Research Center, Linyi Natural Resources and Planning Bureau, Linyi 276000, China; liuyihuichn@whu.edu.cn⁴ Shandong Institute of Land Surveying and Mapping, Jinan 250102, China⁵ Shandong Provincial Territorial Spatial Ecological Restoration Center, Jinan 250014, China⁶ Key Laboratory of Environmental Change and Natural Disaster of Ministry of Education, Beijing Normal University, Beijing 100875, China

* Correspondence: xuxibo@stu.sdu.edu.cn; Tel.: +86-077-2944-3069

Highlights

What are the main findings?

- An improved spectral estimation and mapping framework for soil cadmium (Cd) was developed using geospatial information-guided machine learning (GIGS) and principal component spectra.
- Compared to commonly used models (e.g., XGBoost, RF, SVR), the GIGS model achieved a maximum performance improvement of approximately 27.87% and a minimum improvement of approximately 16.42% (with reference to the R^2 value).

What are the implications of the main finding?

- GIGS-based spectral estimation framework provides an accurate and reliable tool for proximal or satellite estimation of soil Cd, thereby supporting cost-effective soil pollution management and environmental sustainability worldwide.

Abstract

Integrating machine learning with spectral data provides an effective and cost-efficient scheme for estimating cadmium (Cd) contents in soils compared with labor-intensive laboratory analyses. However, conventional machine learning-based spectral estimation methods often yield unsatisfactory performance because they rely on an unrealistic assumption that the functional relationships across geographically distinct subregions are homogeneous, thereby reducing the accuracy and stability of soil Cd estimation. To address this issue, a geospatial information-guided XGBoost (GIGS) model was developed, in which a spatial weighting module was incorporated into the XGBoost framework to account for spatial heterogeneity in functional relationships among sampling locations. The spatial heterogeneity of each soil spectral sample was quantified and assigned a corresponding spatial weight, which was subsequently incorporated during model calibration. The optimal principal component spectra (PCS) derived from spectral data were used as model inputs, with measured soil Cd content as the dependent variable, thereby establishing a robust spectral estimation model for soil Cd. Results indicated that the GIGS model exhibited satisfactory performance in the spectral estimation of soil Cd content, with R^2 , RMSE, and



Academic Editor: Mohammadmehdi Saberioon

Received: 25 May 2026

Revised: 6 July 2026

Accepted: 9 July 2026

Published: 13 July 2026

Copyright: © 2026 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

RPIQ values of 0.78, 0.04, and 2.02, respectively. Compared to commonly used spectral estimation models (e.g., XGBoost, random forest, support vector regression), the GIGS model achieved a maximum performance improvement of approximately 27.87% and a minimum improvement of approximately 16.42% (with reference to the R^2 value). Five PCS of soil spectral data were extracted as predictors for the model. PCS-1 was found to be closely associated with iron oxides, while PCS-2 primarily reflects the spectral characteristics of clay minerals. The spectral bands at 600 nm and 815 nm contribute most strongly to PCS-3, which is linked to soil organic matter and indirectly reflects the soil Cd status. PCS-4 and PCS-5 represent mixed spectral information derived from materials associated with soil Cd. An integrated framework combining the GIGS model and PCS data developed in this study provides an accurate and reliable tool for spectral estimation and mapping of soil Cd, thereby supporting cost-effective soil management and environmental sustainability worldwide.

Keywords: soil potentially toxic elements; spectral estimation; digital mapping; spatial heterogeneity; machine learning

1. Introduction

Soil is a vital component of the Earth's land ecosystem and provides the foundation for human life and production activities [1]. Approximately 95% of the food humans consume comes from soil, while only 5% comes from aquatic sources and other sources [2]. With intensive industrial and agricultural activities, large amounts of pollutants are released into the soil environment. Potentially toxic elements (PTEs) are naturally occurring chemical substances in soils. When combined with large external inputs, they have become some of the most significant pollutants [3–5]. Hou et al. [6] reported that PTE concentrations exceed risk thresholds on about 14% to 17% of arable land worldwide. According to the Food and Agriculture Organization of the United Nations (FAO), by 2050, nearly 90% of global soil resources may be at risk of degradation due to soil erosion, improper fertilizer and pesticide use, and industrial pollution [7].

Cadmium (Cd) is one of the most common PTEs found in soils, with about 19% of soils exceeding the soil quality standard according to a nationwide soil survey in China [8,9]. Cd in soil comes from both natural and anthropogenic sources. Cd is naturally present in bedrock, which acts as the parent material for soil formation [10]. During the processes of bedrock weathering and soil development, Cd gradually moves into the soil through various transformation mechanisms. Anthropogenic sources of Cd in soil mainly include agricultural, domestic, and industrial activities [11]. Facilities and products produced by human activities often contain higher levels of Cd, which can greatly increase the possibility of Cd accumulation in the soil environment [12]. Cd is non-essential and harmful to the human body; exposure can easily cause cancer, genetic mutations, and other health hazards [13]. The accumulation of Cd in soil is considered a significant threat to ecosystems, food safety, and public health [14,15]. Therefore, establishing an accurate and efficient method to detect Cd content in soil would support health risk assessments and pollution control efforts [16].

Proximal or satellite spectra have shown great potential for rapid, efficient, and non-destructive monitoring of soil Cd. The traditional approach for monitoring soil Cd involves soil sampling combined with laboratory chemical analysis [3]. This method is simple and requires low technical expertise; however, it is costly, time-consuming, and inefficient, making it unsuitable for large-scale or repeated monitoring of soil Cd status [17].

Laboratory analyses, such as atomic absorption spectroscopy and inductively coupled plasma mass spectrometry, also generate experimental waste that may cause environmental pollution if not properly managed [18]. In contrast, proximal or satellite spectral sensors can detect soil properties and subtle variations using hundreds of continuous, narrow spectral bands, making it an attractive method for analyzing soil components [3,19]. When combined with data-driven machine learning algorithms, spectral models can effectively extract soil property information and accurately estimate the soil Cd [20,21].

Establishing a functional relationship between the soil Cd contents and spectral data using data-driven machine learning methods is a key step in spectral estimation modeling of soil Cd [22,23]. Data-driven machine learning methods are effective for identifying nonlinear relationships within spectral data and establishing spectral estimation models for soil Cd [24–26]. Compared to traditional spectral estimation models based on empirical-based linear relationships, data-driven machine learning models offer higher accuracy and reliability in the spectral estimation of Cd contents in the soil [27]. The spectral features for the soil Cd are selected from spectral data as independent variables, while the measured Cd contents in soil are used as the dependent variable. The spectral estimation model of soil Cd was constructed after model training, parameter optimization, and validation.

The key assumption for the development of data-driven spectral estimation models of soil Cd is that the soil environment within a target geographic region is homogeneous. Specifically, this assumption presumes that the relationship coefficients between soil Cd and spectral data are consistent across different locations within the target geographical area, thereby allowing data-driven spectral estimation models to effectively identify functional relationships within the spectral data [28–30].

Although such an assumption simplifies the identification of empirical-based relationships from the spectral data and shortens the training time of the model, ignoring the spatial heterogeneity may lead to a spectral estimation model of soil Cd with poor accuracy and low generalization. Under the combined influence of human activities (such as agricultural and industrial activities) and natural environmental factors (winds and rainfall), soil ecosystems within a region exhibit significant spatial heterogeneity. This soil heterogeneity is reflected in differences in soil structure, composition, and habitat types; the influence of soil Cd contamination on spectral data varies across different geographic locations [30,31].

However, the assumption that spectral relationships remain unchanged and that geographical regions are homogeneous in spectral modeling contradicts the actual conditions of regional soil environments [29,32]. As a result, data-driven spectral estimation models often underestimate high soil Cd contamination, which reduces the models' accuracy and reliability.

Therefore, this study proposes a geospatial information-guided machine learning (GIGS) framework to account for spatial heterogeneity in functional relationships among sampling locations, and to enhance the spectral estimation of soil Cd. Firstly, a spatial weighting module was incorporated into the XGBoost model to analyze the spatial heterogeneity of soil spectral samples and assign a spatial weight to each sample during model calibration. This approach eliminates the influence of spatial heterogeneity on the fitting of empirical relationships in spectral modeling. Furthermore, the GIGS model is able to integrate results from both local and global modeling modes, which contributes to further improving the spectral estimation of soil Cd.

This study aims to: (1) develop a GIGS model for estimating soil Cd content; (2) derive principal component spectral (PCS) data to characterize soil Cd; and (3) evaluate the performance of the GIGS model.

2. Materials and Methods

2.1. Study Areas

The study area was a sewage-irrigated region located on the coastal plain north of Longkou City, Shandong Province, China (Figure 1). Its geographical coordinates are (37°34'35"N~37°44'12"N, 120°20'56"E~120°40'47"E), which extends for totally 527.85 km². It belongs to the temperate monsoon climate; the monsoon is more obvious, the four seasons are distinct, the average annual temperature is about 12 °C, and the annual precipitation is about 600 mm~700 mm. The soil types in the study area cover tidal soil, brown soil, and mortar soil, with pH ranging from 6.5 to 7.5 [33]. The study area is located in the Longkou Coal mine area, which is rich in coal resources. Coal mining waste water and coal gangue waste residue caused a serious threat to the soil environment [3]. The residue and waste liquid produced by metal smelters and thermal power plants also aggravated the enrichment of heavy metals in the soil [34,35].

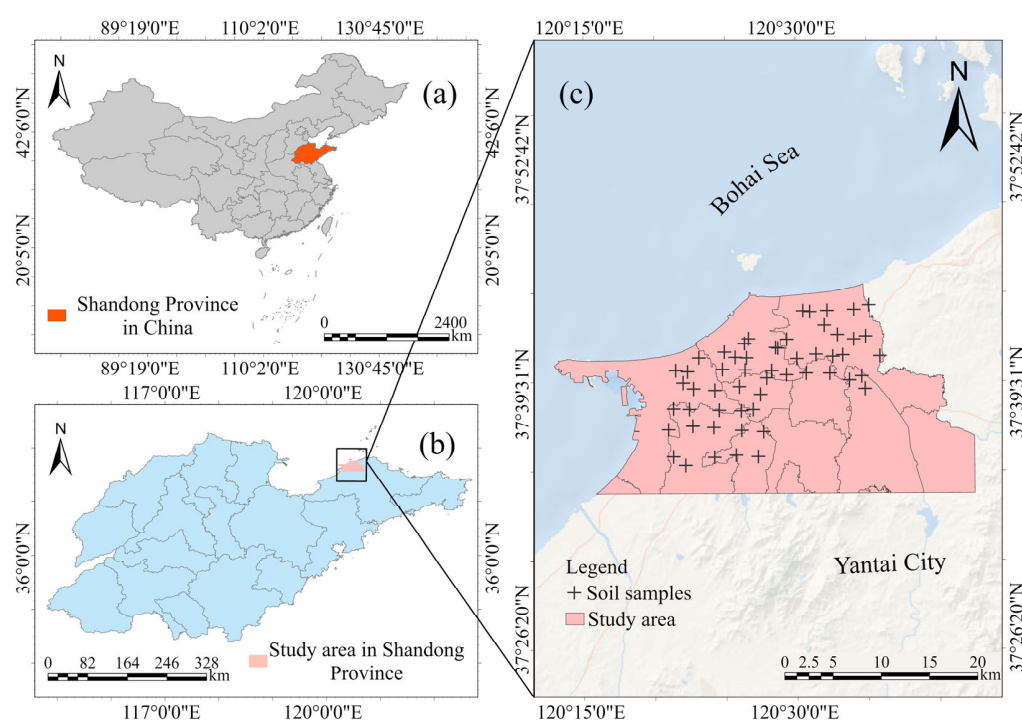


Figure 1. Geographical Locations of the study area and soil samplings: (a) Shandong Province in China; (b) Study area in Shandong Province; (c) Details of the study area and soil sampling.

2.2. Data Collection

2.2.1. Soil Sample Collection

After considering the soil types, land use, and sewage irrigation conditions in the study area, the sampling locations ($n = 67$) were predefined on a digital map using ArcGIS 10.8 software (Esri Inc., Redlands, CA, USA) (Figure 1). At each sampling site, the five-point sampling method was used to collect approximately 1 kg of soil from a depth of 0 to 20 cm. Soil samples were collected from five sites using a shovel. The collected samples were thoroughly mixed to form a composite sample, which was then placed in a clean self-sealing bag. The latitude, longitude, and elevation of each sampling point were determined using a handheld GPS. Additionally, the environmental conditions such as land use and temperature of the sampling locations were recorded with a smart camera. The field survey and sample collection were completed in June 2022.

2.2.2. Soil Spectral Measurement and Chemical Analysis

Soil pretreatment is crucial for the laboratory chemical analysis of soil samples. In the laboratory, the soils were air-dried under natural conditions, and gravel, as well as animal and plant residues, were removed. The soils were then ground and passed through a 100-mesh nylon sieve. Afterward, the soil samples were divided into two parts: one for chemical analysis and the other for spectral measurement.

The ASD FieldSpec-4 portable spectrometer (ASD Inc., Boulder, CO, USA) was used to measure the spectral reflectance of soil samples. The ASD FieldSpec-4 covers a spectral range of 350–2500 nm, with spectral resolutions of 1.4 nm (350–1100 nm) and 2 nm (1000–2500 nm). Soil spectral measurements were conducted in an adjustable light-controlled darkroom. Two 50 W halogen lamps, positioned 30 cm above the sample, were used as the light source, with a lighting angle of 15° and a probe distance of 20 cm.

Before measurement, the spectrometer was preheated for 30 min, and whiteboard calibration was performed to minimize the impact of instrument errors on data acquisition. Whiteboard calibration was repeated several times during the measurement process to reduce noise effects. To avoid random errors, each soil sample was measured five times, and the spectral ranges with a low signal-to-noise ratio were removed [3]. Abnormal spectra were identified, and the average of the remaining spectra was taken as the spectral data for the soil sample.

The inductively coupled plasma mass spectrometer (ICP-MS) method was used to determine the Cd contents in the soil samples [36]. Blank samples and standard soil samples (GBW 07449) were incorporated into the laboratory experiments for soil sample analysis [37]. Ten percent of the samples were randomly selected as parallel samples, with the relative standard deviation (RSD) of their measurements being less than 5%. The recovery rates of the standard samples ranged from 95% to 105%, demonstrating the reliability of the laboratory analysis for Cd in soil samples.

2.3. Principal Component Spectral Data

The direct spectral signatures of soil Cd are difficult to capture due to its low concentration and environmental interference. Therefore, spectrally active substances associated with Cd, such as soil organic matter, iron oxides, and clay minerals, are commonly used as proxies for soil Cd content. In this study, principal component transformation was applied to hyperspectral data for dimensionality reduction and feature extraction. The resulting principal component spectral (PCS) data, which represent spectral information related to Cd-associated materials, were used as input variables for spectral modeling, thereby reducing model convergence time and improving model robustness.

2.3.1. First-Order Derivative

Spectral transformation aimed to highlight the spectral signals of soil Cd and eliminate the noise information generated by environmental factors and instrument errors [38,39]. In this study, the Savitzky–Golay (SG) and First-order derivative (FOD) were used to transform the soil spectra data [40,41]. The SG transformation is a technique used to reduce random noise in soil spectral data. The SG transformation applies a k th-order polynomial fitting within a specified window width to the data. The expression for the SG transformation is as follows:

$$R_i = \frac{h_n}{H} \sum_{n=-w}^{+w} R_{k+n} \quad (1)$$

where R_i represents the i th spectral band, w is the filter window width, H is the normalization coefficient, and h_n is the smoothing coefficient, which is obtained by least-squares polynomial fitting.

First-order derivative (FOD) transformation is an effective technique to enhance features in soil spectra, including inflection points of the spectral curve and positions of reflectance extrema [42]. These features are associated with the Cd content in the soil. The FOD can be expressed as:

$$\text{FOD}(\lambda_i) = \frac{R(\lambda_{i+1}) - R(\lambda_{i-1})}{2\Delta\lambda} \quad (2)$$

where $R(\lambda_{i+2})$, $R(\lambda_{i+1})$, and $R(\lambda_{i-2})$ indicate the reflectance values at the spectral bands of the $(i+2)$ th, i th, and $(i-2)$ th, respectively. $\Delta\lambda$ represents the wavelength interval between the i th spectral band and the $(i+1)$ th spectral band.

2.3.2. Principal Component Transformation

Principal component transformation is a dimensionality reduction method that can extract the effective information of soil spectral data through information compression by using the orthogonal transformation [43,44]. Soil spectral data included a large number of spectral bands, and collinearity and redundancy existed. Principal component analysis can effectively reduce the data dimension, extract key information in the soil spectral data, and support the GIGS model construction. Principal component transformation was expressed as follows:

$$Z = \frac{X - \bar{X}}{d} \quad (3)$$

where X is the spectral reflectance, $\bar{X}$ is the average of the spectral reflectance, and d represents the standard deviation.

$$\text{cov}(x_i, y_j) = \frac{\sum_{k=1}^n (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{n - 1} \quad (4)$$

where k indicates the k -th soil sample, i and j denote the i -th and j -th spectral bands, respectively, X_{ki} is the reflectance of the i -th band of the k -th soil sample, and n is the number of soil samples.

$$c = \frac{\lambda_k}{\sum_{i=1}^p \lambda_i} \quad (5)$$

where c represents the cumulative variance contribution rate of the principal component spectrum, λ represents the eigenvector, and i is the number of eigenvectors. Principal component calculations were performed using the Python software (Version 3.14).

2.4. GIGS Model

The GIGS model is an improved spectral estimation model for soil Cd that was developed using the XGBoost modeling framework (Version 2.1.0) and offers advantages in soil properties estimation in geographic environments with significant spatial heterogeneity [45,46]. The GIGS model integrates the results of two sub-models under global and local modes, and incorporates a spatial weighting module, thereby improving the robustness of soil Cd spectral estimation [47,48]. The framework for the GIGS model is illustrated in Figure 2.

We assume that a soil spectral dataset consisting of n samples was collected, where y_i denotes the target variable and x_i represents the explanatory variable at location i . The GIGS model under the global mode can be expressed as follows:

$$\hat{y}_g = \sum_{k=1}^K f_k(x_i), f_k \in F \quad (6)$$

where $\hat{y}_g$ indicates the estimated value at location i , K is the number of decision trees, and f_k is a function belonging to the function space F ; F represents the set of all possible decision trees.

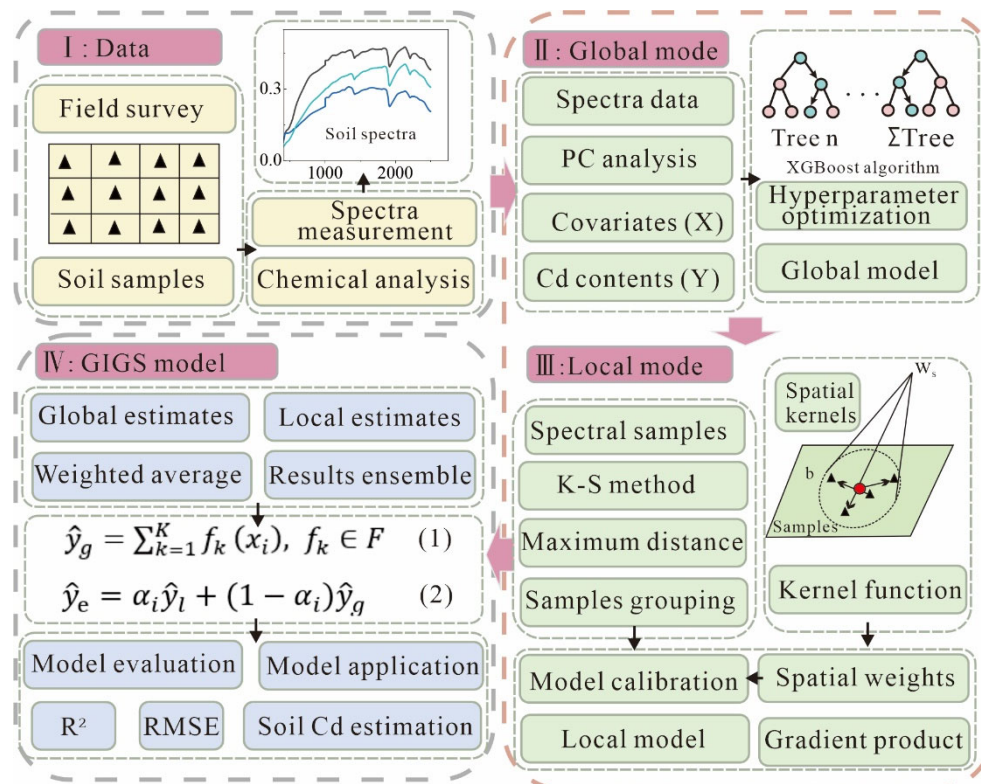


Figure 2. Flowchart for the construction of the geospatial information-guided spectral (GIGS) model.

The GIGS model in the local mode consists of three steps. First, the distances between all spectral samples are calculated using the Kennard–Stone algorithm, and two sample subsets are partitioned based on the principle of maximizing spectral distance [49]. Subsequently, a neighborhood is constructed within each sample subset. The Gaussian function is defined as the kernel function, and spatial kernels are used to calculate the spatial weights (w_s) for each sample subset [47].

$$ws_{ij} = \exp\left(-\frac{d_{ij}^2}{b^2}\right) \quad (7)$$

where d_{ij} represents the Euclidean distance between the i th and j th samples with different geographic locations, and b is the threshold defined for the distance.

Spatial weights function in the GIGS model are used to assign the spatial weight for the spectral samples. By adopting a gradient product approach, the conventional weights obtained from the XGBoost algorithm are replaced with spatial weights during the model calibration [50]. The spatial weights are multiplied by the first- and second-order gradients, assigning a weight to the spectral samples. In addition, the GIGS model can also be calibrated by the soil spectral samples without the spatial weights.

In the end, the estimation results from the local mode and global mode in the GIGS model were integrated by using a weighted averaging approach, the ensemble model as shown in the following equation:

$$\hat{y}_e = \begin{bmatrix} \hat{y}_g & \alpha_i = 0 \\ \alpha_i \hat{y}_l + (1 - \alpha_i) \hat{y}_g & \alpha_i \in (0, 1) \\ \hat{y}_l & \alpha_i = 1 \end{bmatrix} \quad (8)$$

where $\hat{y}_e$ indicates the estimated value of soil Cd from the GIGS model at location i . α is a hyperparameter of the GIGS model, related to the spatial heterogeneity of the soil spectral samples, and is used to guide the model. The α value ranges from 0 to 1 and is optimized by the residuals of the global ($\hat{y}_g$) and local sub-models ($\hat{y}_l$). The estimation of the ensemble model is the weighted average of the prediction results from the local and global sub-models at location i . The case where α equals zero is a special scenario in which the estimation at location i is computed solely by the global mode of the GIGS model, without considering the local mode.

During the calibration of the GIGS model, the hyperparameters in the objective function, loss function, and regularization term for each decision tree are optimized by grid search and cross-validation, thereby improving the accuracy of the GIGS model and avoiding overfitting [46,50,51]. The GIGS model was implemented using Python software (Version 3.14).

2.5. Model Calibration and Evaluation

Model performance was evaluated using three indices: coefficient of determination (R^2), root mean square error (RMSE), and ratio of performance to quartiles (RPIQ) [52,53]. A well-performing model is characterized by lower RMSE values and higher R^2 and RPIQ values [54]. The collected soil samples ($n = 67$) were divided into the calibration set ($n = 50$) and the validation set ($n = 17$) based on the Kennard–Stone algorithm. The calibration set was used to train the GIGS model, while the validation set was used for independent evaluation of model performance and did not participate in model calibration.

In addition, several commonly used machine learning algorithms were applied to construct spectral estimation models of soil Cd, including extreme gradient boosting (XGBoost), random forest (RF), and support vector regression (SVR) [55,56]. Five-fold cross-validation and grid search were used to tune the hyperparameters to achieve the best performance of the above models. The accuracy of these algorithms was evaluated to provide a better assessment of the GIGS model's performance.

3. Result and Analysis

3.1. Descriptive Statistics Characteristics of the Cd Contents in the Soil Samples

The descriptive statistical characteristics of Cd contents in soils of sewage irrigation areas are shown in Figure 3. The soil Cd contents range from 0.14 to 0.87 mg kg^{−1}, with a mean of 0.33 mg kg^{−1} and a standard deviation of 0.12 mg kg^{−1}. The coefficient of variation is 35.65%. This coefficient, which is a dimensionless measure of standard deviation, reflects the variability of Cd concentrations among soil samples [10,57]. The coefficient of variation is less than 36% and higher than 15%, indicating that Cd exhibits moderate spatial heterogeneity in the study area.

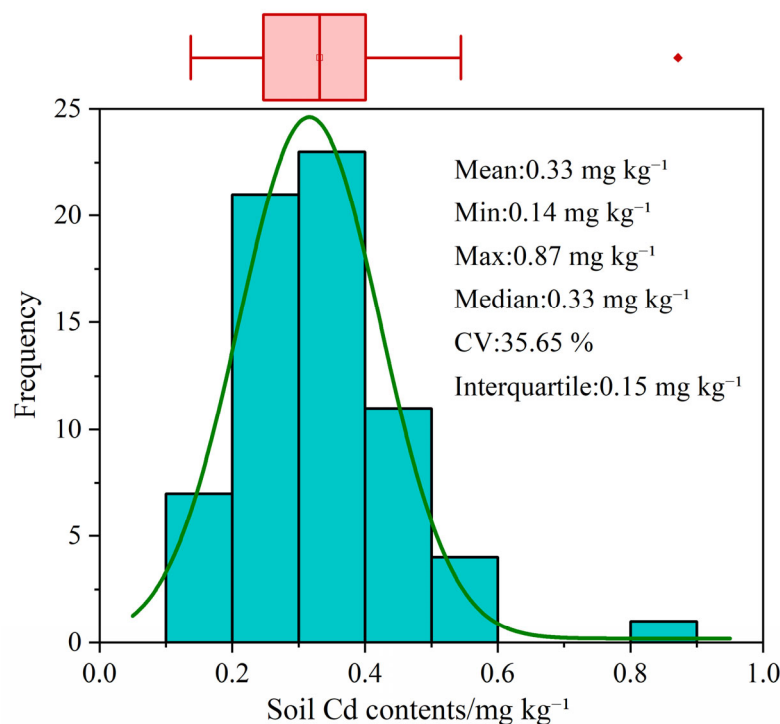


Figure 3. Descriptive statistical characteristics of the Cd contents in the soil samples in the sewage irrigation areas, China/(mg kg⁻¹).

According to the “Agricultural Land Pollution Control Standard in China” (GB 15618-2018) [58,59], the threshold for Cd pollution is 0.30 mg kg⁻¹. The mean Cd content (0.331 mg kg⁻¹) in the sewage irrigation areas exceeds this threshold, indicating a notable level of Cd pollution in the study area. One of the main sources of soil Cd is the untreated wastewater, exhaust gas, and solid waste discharged from industrial and coal mining activities. These pollutants are transported to rivers and farmlands through hydrological processes and atmospheric transport, leading to soil Cd pollution. A comprehensive analysis of the statistical characteristics of soil Cd not only provides a basis for the construction and optimization of the GIGS model, but also supports agricultural practices and precise environmental management.

3.2. Soil Spectral Analysis

The spectral reflectance of soil samples with different Cd contents shows clear differences in the Vis-NIR region (Figure 4a). Overall, the spectral curves (in the wavelength range of 400–2500 nm) of all soil samples increase, but the rate of increase differs across spectral ranges. Between 500–1000 nm, reflectance rises rapidly, while from 1000–2500 nm, it increases more slowly. Distinct absorption features around 1400 nm and 1900 nm are mainly related to soil moisture. The absorption feature near 2200 nm, as noted by [60,61], is mainly associated with the clay content in the soil. Figure 4b shows the average spectral curves with the FOD transformation. After FOD transformation, the spectral features of soils become more pronounced, local characteristics are enhanced, and small spectral changes are amplified. In addition, Figure 4b also reveals a negative correlation between soil spectral reflectance and Cd content. The spectral curves of soils with different Cd contents are more clearly separated, laying a solid foundation for building accurate spectral models to estimate soil Cd content.

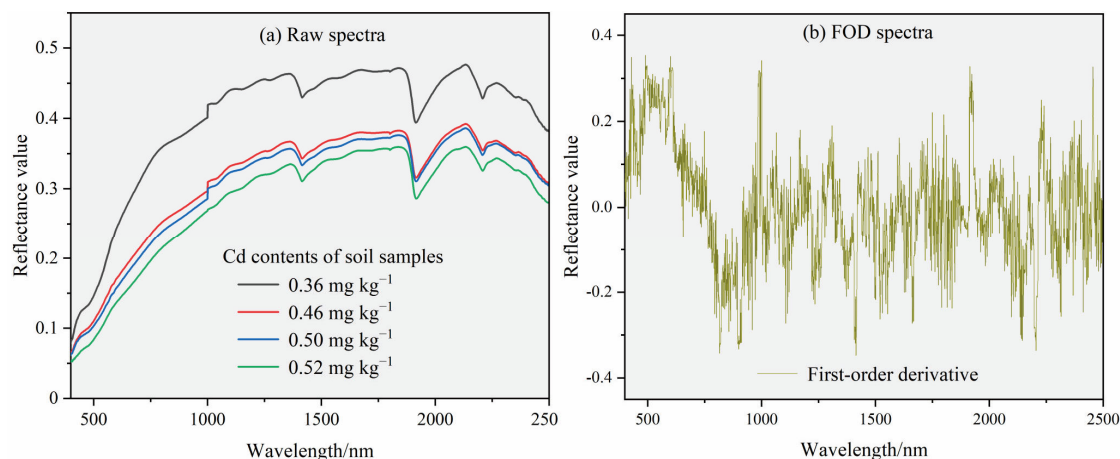


Figure 4. Spectral curves for soil with different Cd contents and the first-order derivative (FOD) spectra data.

3.3. Selecting the PCS from the Spectral Data as Covariates

PCS analysis was performed on the soil spectral data to extract the key information and remove redundant spectral data. The first ten principal components and their eigenvalues are presented in Figure 5. PCS-1 explained 45.85% of the total variance in the spectral data, while PCS-2 and PCS-3 accounted for 23.36% and 10.56% of the total variance, respectively. In addition, to improve the computational efficiency of the spectral model and prevent overfitting, the five PCS were selected as covariates for spectral modeling.

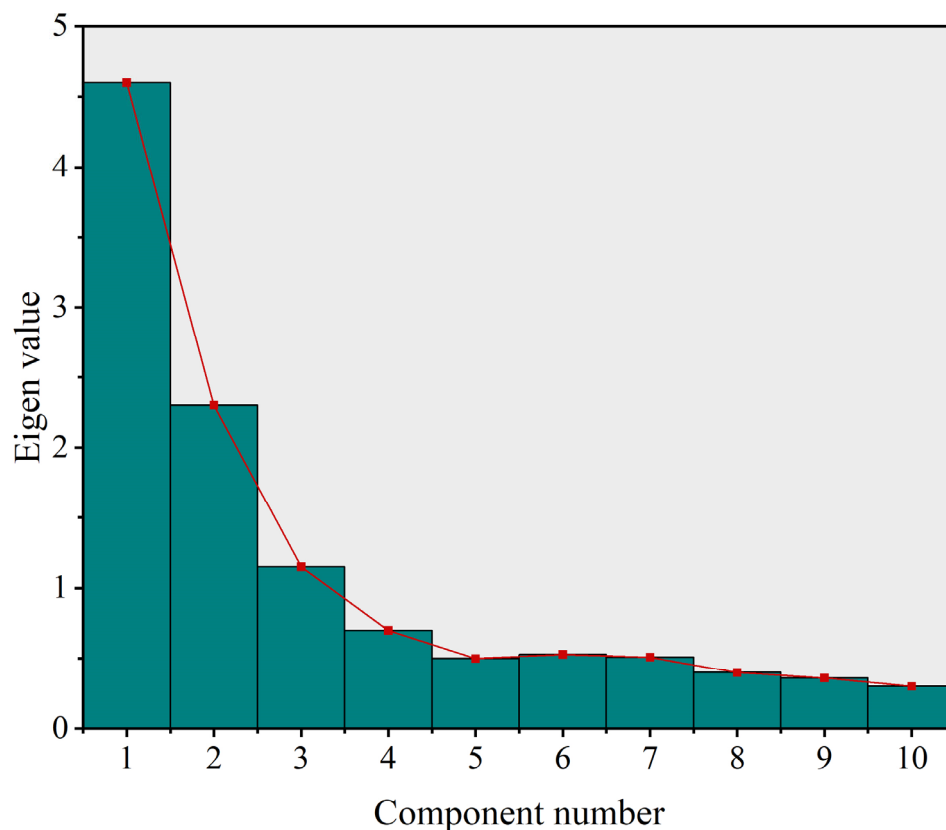


Figure 5. Key information and the principal component spectra (PCS) extracted from the soil spectra data.

3.4. Model Comparison and Evaluation

Five PCS were used as input variables, and the measured Cd content in the soil samples was taken as the dependent variable. Soil Cd spectral estimation models were

developed using GIGS, XGBoost, RF, and SVR algorithms. The results of these models are presented in Figure 6. Detailed accuracy metrics can be found in Table S1. XGBoost, RF, and SVR models exhibited moderate performance, with RPIQ values of 1.35, 1.01, and 1.15, and R^2 values of 0.67, 0.61, and 0.63, respectively. In contrast, the GIGS model achieved satisfactory performance, with the precision indices of R^2 , RMSE, and RPIQ of 0.78, 0.04, and 2.02, respectively. Compared to the other models, the performance of the GIGS model improved by 16.42%, 27.87%, and 23.81% (with reference to the R^2 value), respectively.

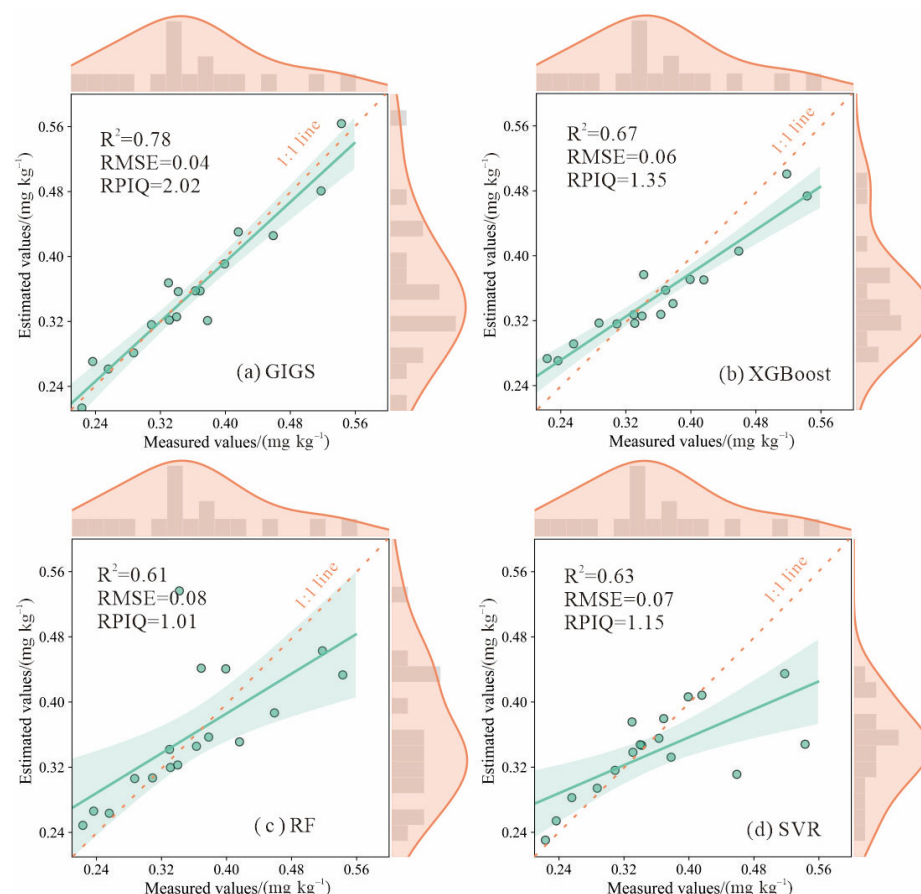


Figure 6. Scatter plots of measured versus estimated Cd contents in soils, and the accuracy indices for the geospatial information-guided spectral (GIGS), extreme gradient boosting (XGBoost), random forest (RF), and support vector regression (SVR) models.

4. Discussion

4.1. Mechanism Exploration of the Principal Component Spectral Data of Soil Cd

In this study, a spatial weighting module was added to the GIGS model to analyze the spatial heterogeneity of soil spectral samples. Each sample was given a spatial weight during model calibration. The estimation function captured by the GIGS model effectively explains the spatial heterogeneity of the spectral samples. Moreover, the GIGS model combines both local and global models for the spectral estimation of soil Cd, which improves the accuracy of spectroscopic estimation in sewage irrigation areas.

The information from spectrally active bands related to soil Cd was effectively interpreted by the extracted PCS. The loading values of these bands on the PCS are shown in Table 1. Scores of spectral samples on the principal components in the projection space are indicated in Figure 7. The absolute loading values at 429 nm, 490 nm, and 1000 nm were significant on PCS-1, indicating that PCS-1 captures most of the information for these spectral bands. The 490 nm and 1000 nm bands are absorption peaks that have been reported to be closely related to soil iron oxides [62]. Therefore, PCS-1 can be regarded

as an indirect spectral indicator of soil Cd associated with iron oxides [63]. The loading values at 900 nm, 1415 nm, 1915 nm, and 2205 nm were higher than those of other spectrally active bands on PCS-2, and the spectral information contained in these bands showed the strongest correlation with PCS-2. The bands at 1415 nm and 1915 nm correspond to the combined spectral features of moisture-related (-OH) stretching vibrations and Al-OH bending vibrations in soil clay minerals [22,60,64]. The 2205 nm band is an absorption peak of soil clay minerals, and the peak near 900 nm is caused by Fe^{3+} in the soil, indicating that PCS-2 mainly reflects the information of soil Fe^{3+} , moisture, and clay minerals [65]. The spectral information of PCS-3 was mainly derived from the bands at 600 nm and 815 nm. The 600 nm band is a typical reflection peak for soil organic matter, and the region near 815 nm is a sensitive band for organic matter response [66,67]. Thus, the indirect spectral indicators for soil Cd in PCS-3 represent the spectral information of soil organic matter [31,68]. Overall, the extracted PCS represent the comprehensive spectral information of soil Cd and indirectly characterize Cd in sewage-irrigated soils through indicators such as organic matter, clay, and iron oxides.

Table 1. Loading values of important spectral bands on the top three principal component spectra (PCS).

Spectrally Active Bands	PCS-1	PCS-2	PCS-3
429 nm	−0.711	−0.448	0.361
490 nm	0.758	−0.492	0.201
600 nm	0.661	−0.361	0.792
815 nm	0.410	0.448	0.867
900 nm	0.431	0.766	0.418
1000 nm	−0.937	−0.041	−0.010
1415 nm	0.494	0.778	−0.140
1915 nm	0.341	−0.843	−0.137
2205 nm	−0.074	0.812	0.203
2455 nm	0.694	0.700	−0.103

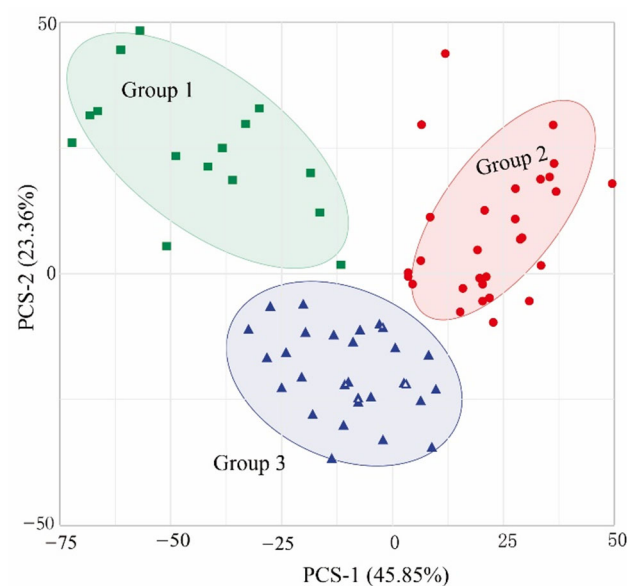


Figure 7. Scores of spectral samples on the principal components in the projection space.

4.2. Performance and Error Analysis of the GIGS Model

The sample set division is important for calibrating the GIGS model using both local and global modeling approaches. Moosavi et al. [69] and Shi et al. [70] found that local

models can perform better than global models in estimating soil properties from spectra, because they take into account the spatial heterogeneity of soil spectral samples. Choosing a suitable sample set is key to achieving the best performance of the local model in the GIGS model [48,71,72]. In this study, the Kennard–Stone algorithm was used to divide the calibration sample set into two equally sized subsets. Although this was the best choice for the current study, if the total number of samples, spatial heterogeneity, or soil properties change, the way the sample set is divided should be reassessed.

The GIGS model can recognize the spatial heterogeneity in soil spectral samples, but it does not model the spatial autocorrelation of samples within specific geographic environments. Due to spatial heterogeneity, the functional relationship and those weight coefficients between spectral data and soil Cd content vary across different locations [29,30,32]. The GIGS model addresses this negative effect from the spatial heterogeneity by using spatial weighting in the modeling process. Furthermore, the GIGS model did not achieve perfect accuracy (with R^2 and RPIQ values of 0.78 and 2.02, respectively). This shows that the spatial relationships or dependencies among soil spectral samples within a given region are not fully represented, which impacts the performance of the model. At present, the soil environment in sewage irrigation areas is complex and highly variable [73]. The GIGS model offers a way to reduce the negative influence of spatial heterogeneity on soil spectral estimation.

Both conventional machine learning models and the GIGS-based spectral estimation model were run for 50 iterations, with the resulting standard deviations of soil Cd estimates serving as indicators of model accuracy and uncertainty. The distribution of standard deviations across all soil samples is illustrated in Figure 8. Among the four models evaluated, GIGS (Figure 8a) achieved the lowest mean standard deviation (0.045), indicating the most stable and consistent estimation performance. XGBoost (Figure 8b) and RF (Figure 8c) ranked second and third, with mean standard deviations of 0.059 and 0.065, respectively, while SVR (Figure 8d) exhibited the largest variability, with a mean standard deviation of 0.071. To further investigate spatial performance, the study area was partitioned into four subregions (E, W, N, and S) based on differences in soil environmental conditions and anthropogenic disturbance. Subregion E displayed the highest coefficient of variation in soil Cd content (44.72%), substantially exceeding those of subregions W (36.13%), N (35.57%), and S (33.17%), thus reflecting significant spatial heterogeneity. GIGS maintains a relatively low standard deviation (0.055) in subregion E, while all three conventional machine learning models exhibit values exceeding 0.06. These findings highlight the superior performance of the GIGS model for spectral estimation of soil Cd, particularly in areas characterized by strong spatial heterogeneity.

4.3. Transferability and Limitations Analysis

To evaluate the transferability and robustness of the GIGS model, two case study areas, Changqing County (CQC), China, and Thessaly and Central Greece (TCG), Greece (see Figure 9), were selected for validation. The CQC is located in eastern China, with geographic coordinates ranging from 36°14' to 36°41'N and 116°30' to 117°4'E, covering an area of approximately 1180 km². It belongs to a warm temperate continental monsoon climate zone. The region lies on the alluvial plain of the lower Yellow River and is characterized by three main soil types: brown soil, loam, and alluvial soil. The TCG is situated in the central plain of Greece, covering an area of about 1315 km². Its geographic coordinates range from 39°23' to 39°45'N and 22°10' to 22°47'E. The region has a typical Mediterranean climate. The main soil types include brown soil, terra rossa, and vertisols. The dominant crops in this area are wheat, lentils, and cotton.

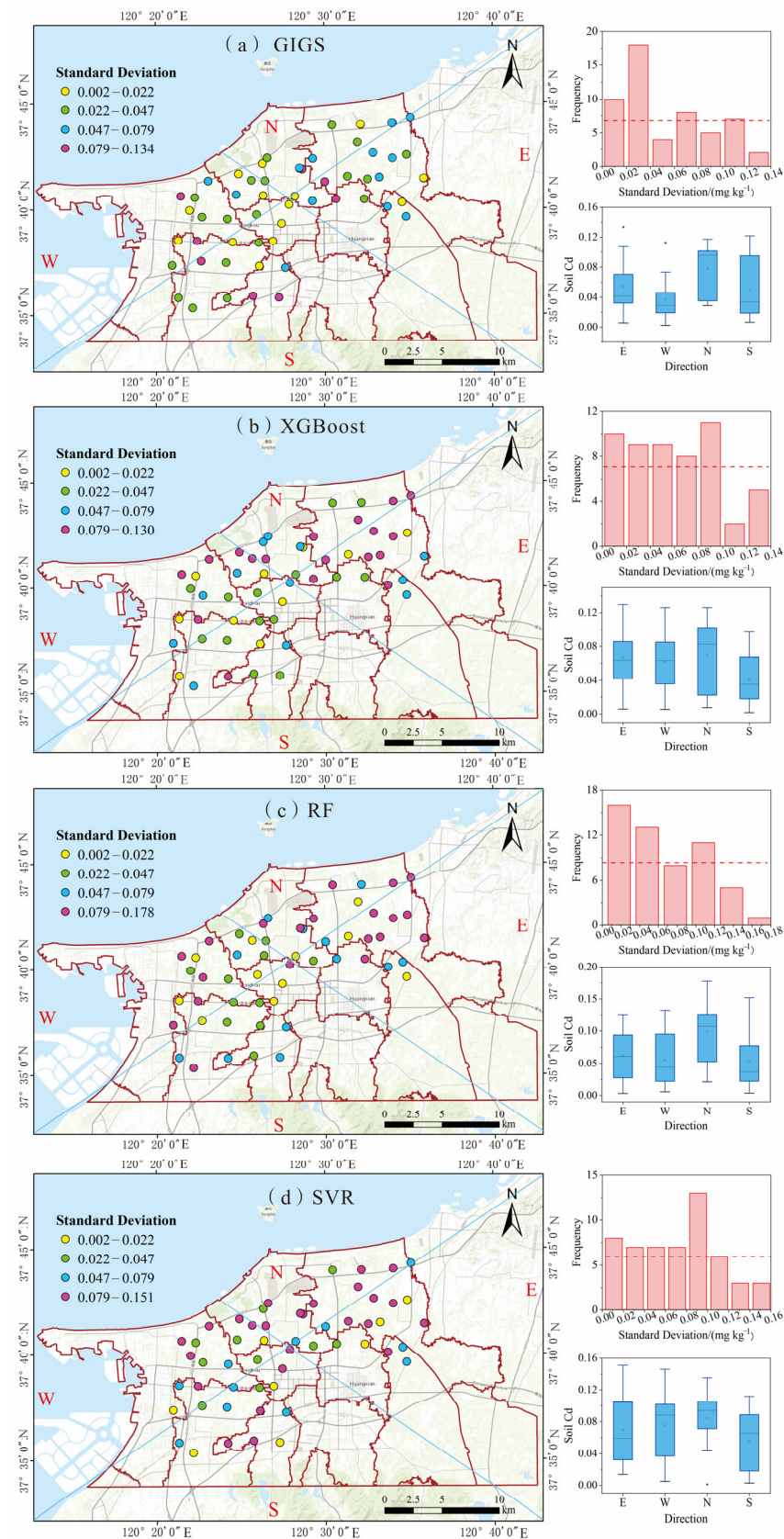


Figure 8. Standard deviation of soil Cd spectral estimation using the GIGS model in a sewage irrigation area of China: The solid blue lines represent the boundaries of the four subregions within the study area; the solid red lines indicate the boundaries of the towns within the study area; and the dashed red lines represent the mean frequency of standard deviations across different ranges.

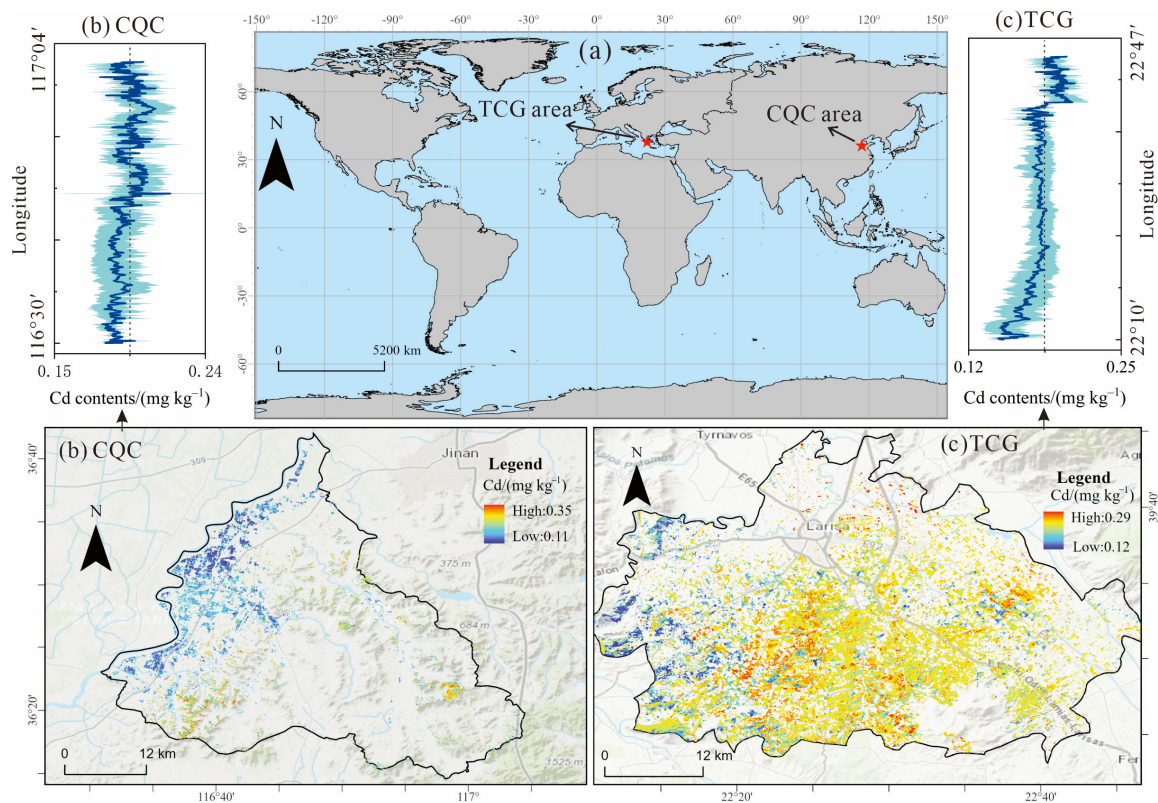


Figure 9. Transferability of the GIGS model in two case study areas, Changqing County (CQC), China, and Thessaly and Central Greece (TCG), Greece: (a) The geographic locations of the CQC and TCG areas in the world; (b) Estimated map of soil Cd in the CQC area; (c) Estimated map of soil Cd in the TCG area.

The soil samples and corresponding spectral data for both the CQC and TCG areas were obtained from publicly available datasets [53,74]. A total of 304 soil samples were collected in the CQC area. Near-synchronous Landsat 8 OLI imagery (22 April 2020) was acquired from the Google Earth Engine platform to derive PCS data corresponding to the soil samples. Detailed information about the dataset can be found in Xu et al. [53]. Similarly, soil samples for the TCG area were obtained from Tóth et al. [74], along with near-synchronous Landsat 8 OLI imagery (26 July 2013). The preprocessing of soil samples and spectral data is described in Section 2, which supports a comprehensive and reliable evaluation of the GIGS model framework.

The estimated soil Cd and its spatial distribution are shown in Figure 9. The GIGS model achieved acceptable accuracy in both the CQC area and TCG area for spectral estimation of soil Cd. The R^2 and RPIQ values were 0.59 and 1.00 for the CQC area, and 0.64 and 1.17 for the TCG area. As shown in Figure 9b, areas with high soil Cd content are mainly distributed in the eastern part of the CQC area, which may be related to the parent material of the soil [10,16]. Figure 9b also indicates that hotspots of soil Cd contamination in the TCG area are mainly located in the eastern region, which are associated with agricultural activities and parent material [35,74]. These results provide evidence for the reliability and transferability of the GIGS model.

The GIGS model is combined with Sentinel-2 MSI imagery to generate a distribution map of soil Cd in the study area (Figure 1). The Cd content in the soils in the northern part of the study area is higher than that in the southern part, which is related to industrial activities and coal mining (as indicated in Figure S1). The mapping results show that soil Cd contents range from 0.12 mg kg⁻¹ to 0.79 mg kg⁻¹. This range is basically consistent with the statistical results from ground-based soil samplings (0.14 mg kg⁻¹ to 0.87 mg kg⁻¹).

These findings confirm that the GIGS model framework is effective for proximal estimation of soil Cd and for satellite-based mapping.

Selecting strong spectral features as input variables can improve the performance of the GIGS model. Advanced supervised search methods such as the genetic algorithm and successive projection algorithm have been used to select spectral features from the numerous bands in soil spectral data [75–77]. The selected spectral features are closely related to soil components, which can greatly improve the accuracy of soil spectral estimation models. Although the PCS analysis method used in this study helps extract key spectral information and reduce data redundancy and collinearity, further research is needed to achieve more precise spectral feature extraction and to develop robust spectral feature sets.

Future research should focus on developing new theoretical frameworks for the spectral estimation and monitoring of soil Cd. The GIGS method proposed in this study improves estimation accuracy by accounting for the spatial heterogeneity of soil spectral samples and reducing spatial non-stationarity between spectral data and Cd content. This study represents a meaningful improvement over existing theoretical frameworks for soil Cd spectral estimation [20,78]. Building on this work, new theoretical frameworks should be developed in the context of rapid advances in AI technologies to enable the analysis of spatiotemporal dynamics, driving factors, and future trends of soil Cd [79].

We acknowledge that the GIGS approach has several limitations that should be carefully considered and addressed as a priority. First, the XGBoost model shows stable and strong performance across different application scenarios [50,78]. Considering the balance between efficiency and cost, the GIGS approach was developed based on an XGBoost model framework with a spatial weighting module. In future work, the GIGS approach should be further improved by incorporating advanced AI frameworks, such as transformer-based deep learning and pretrained foundation models [79,80]. In addition, although representative soil samples were collected for calibrating the GIGS model, and both independent validation and cross-validation were used to improve model reliability, limitations such as small sample size, limited validation samples, and the risk of overfitting cannot be avoided and should be addressed in future studies. Finally, the spectral estimation and digital mapping of soil Cd in this study focused on topsoil (0–20 cm). Future work can integrate multi-source remote sensing data and soil process models to obtain PTEs information in deeper soil layers [81].

5. Conclusions

Integrating the PCS data with a GIGS model developed in the study provided a framework for efficiently estimating and mapping soil Cd content and achieved satisfactory performance in a case study. A spatial weighting module from the GIGS model was used to analyze the spatial heterogeneity of each soil spectral sample. Each sample was assigned a spatial weight based on its geographical location, which helped reduce the negative effects of spatial heterogeneity on spectral modeling. The optimal principal component spectra (PCS) derived from spectral data were used as model inputs, with measured soil Cd content as the dependent variable, thereby establishing a robust spectral estimation model for soil Cd. Results indicated that the GIGS model exhibited satisfactory performance in the spectral estimation of soil Cd content, with R^2 , RMSE, and RPIQ values of 0.78, 0.04, and 2.02, respectively. Compared to commonly used spectral estimation models (e.g., XGBoost, RF, SVR), the GIGS model showed an improvement of about 27.87% at the highest and about 16.42% at the lowest (based on R^2). Five PCS of soil spectral data were extracted as predictors for the model. These predictors were closely associated with clay minerals, organic matter, and iron oxides in the soil, thereby indirectly reflecting the state of Cd content. This study provides an efficient spectral estimation and mapping framework for

soil Cd content, which can strongly support soil quality monitoring and environmental sustainability around the world.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/rs18142341/s1>: Table S1: Accuracy indices for the GIGS, XGBoost, RF, and SVR models; Figure S1: Map of soil Cd distribution in the study area.

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

Funding: This work was supported by the Natural Science Foundation of Shandong Province, China (ZR2025QC1023; ZR2024QD268), National Natural Science Foundation of China (42501074), and Key Laboratory of Environmental Change and Natural Disaster of the Ministry of Education, Beijing Normal University (2023-KF-12).

Data Availability Statement: The datasets used in this study can be obtained from the corresponding author upon reasonable request.

Acknowledgments: We gratefully acknowledge the support of the State Key Laboratory of Earth Surface Processes and Disaster Risk Reduction at Beijing Normal University.

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

References

- Chen, M.; Zhang, X.; Jiang, P.; Liu, J.; You, S.; Lv, Y. Advances in heavy metals detoxification, tolerance, accumulation mechanisms, and properties enhancement of *Leersia hexandra* Swartz. *J. Plant Interact.* **2022**, *17*, 766–778. [\[CrossRef\]](#)
- Spanner, J.; Napolitano, G. *Healthy Soils Are the Basis for Healthy Food Production*; FAO: Rome, Italy, 2015; pp. 1–4.
- Wang, Y.; Zhang, X.; Sun, W.; Wang, J.; Ding, S.; Liu, S. Effects of hyperspectral data with different spectral resolutions on the estimation of soil heavy metal content: From ground-based and airborne data to satellite-simulated data. *Sci. Total Environ.* **2022**, *838*, 156129. [\[CrossRef\]](#) [\[PubMed\]](#)
- Guo, Z.; Zhang, Y.; Xu, R.; Xie, H.; Xiao, X.; Peng, C. Contamination vertical distribution and key factors identification of metal(loid)s in site soil from an abandoned Pb/Zn smelter using machine learning. *Sci. Total Environ.* **2023**, *856*, 159264. [\[CrossRef\]](#) [\[PubMed\]](#)
- Yin, G.; Chen, X.; Zhu, H.; Chen, Z.; Su, C.; He, Z.; Qiu, J.; Wang, T. A novel interpolation method to predict soil heavy metals based on a genetic algorithm and neural network model. *Sci. Total Environ.* **2022**, *825*, 153948. [\[CrossRef\]](#) [\[PubMed\]](#)
- Hou, D.; Jia, X.; Wang, L.; McGrath, S.P.; Zhu, Y.-G.; Hu, Q.; Zhao, F.-J.; Bank, M.S.; O'Connor, D.; Nriagu, J. Global soil pollution by toxic metals threatens agriculture and human health. *Science* **2025**, *388*, 316–321. [\[CrossRef\]](#) [\[PubMed\]](#)
- Jeffrey, M.; Achurch, H. Save our soil to save the planet. In *Global Soil Security*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 389–395.
- Suhani, I.; Sahab, S.; Srivastava, V.; Singh, R.P. Impact of cadmium pollution on food safety and human health. *Curr. Opin. Toxicol.* **2021**, *27*, 1–7. [\[CrossRef\]](#)
- Ministry of Ecology and Environment of the People's Republic of China. *National Soil Contamination Survey Report*; Ministry of Ecology and Environment of the People's Republic of China: Beijing, China, 2014.
- Lv, J.; Liu, Y.; Zhang, Z.; Dai, J. Factorial kriging and stepwise regression approach to identify environmental factors influencing spatial multi-scale variability of heavy metals in soils. *J. Hazard. Mater.* **2013**, *261*, 387–397. [\[CrossRef\]](#) [\[PubMed\]](#)
- Meharg, A.A.; Meharg, C. The pedosphere as a sink, source, and record of anthropogenic and natural arsenic atmospheric deposition. *Environ. Sci. Technol.* **2021**, *55*, 7757–7769. [\[CrossRef\]](#) [\[PubMed\]](#)
- Nriagu, J.O.; Pacyna, J.M. Quantitative assessment of worldwide contamination of air, water and soils by trace metals. *Nature* **1988**, *333*, 134–139. [\[CrossRef\]](#) [\[PubMed\]](#)
- Wang, X.; Liu, X.; Yang, X.; Cheng, T.; Yang, Z. Environmental risks for application of iron and steel slags in soils in China: A review. *Pedosphere* **2021**, *31*, 28–42. [\[CrossRef\]](#)
- Liu, S.; Zhao, H.; Wu, K.; Zhang, Z.; Hou, Y.; Chen, T.; Jin, Q. Evaluation of heavy metal distribution characteristics of agricultural soil–rice system in a high geological background area according to the influence index of comprehensive quality (IICQ). *Environ. Sci. Pollut. Res.* **2020**, *27*, 20920–20933. [\[CrossRef\]](#)

15. Jia, Z.; Wang, J.; Li, B.; Li, Y.; Zhou, Y.; Tong, G.; Yan, D.; Zhou, S. An integrated methodology for improving heavy metal risk management in soil-rice system. *J. Clean. Prod.* **2020**, *273*, 122797. [[CrossRef](#)]
16. Li, J.; Li, K.; Jiao, L.; Zang, F.; Li, X.; Yang, Y.; Mao, X.; Tai, X. Contamination, ecological-health risks, and sources of potentially toxic elements in road-dust sediments and soils. *Environ. Geochem. Health* **2023**, *45*, 8169–8186. [[CrossRef](#)] [[PubMed](#)]
17. Rossel, R.V.; Walvoort, D.; McBratney, A.; Janik, L.J.; Skjemstad, J. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma* **2006**, *131*, 59–75. [[CrossRef](#)]
18. Van Acker, T.; Theiner, S.; Bolea-Fernandez, E.; Vanhaecke, F.; Koellensperger, G. Inductively coupled plasma mass spectrometry. *Anal. Chem.* **2023**, *95*, 52–64. [[CrossRef](#)]
19. Dian, Y.; Le, Y.; Fang, S.; Xu, Y.; Yao, C.; Liu, G. Influence of spectral bandwidth and position on chlorophyll content retrieval at leaf and canopy levels. *J. Indian Soc. Remote Sens.* **2016**, *44*, 583–593. [[CrossRef](#)]
20. Wang, Y.; Zou, B.; Chai, L.; Lin, Z.; Feng, H.; Tang, Y.; Tian, R.; Tu, Y.; Zhang, B.; Zou, H. Monitoring of soil heavy metals based on hyperspectral remote sensing: A review. *Earth-Sci. Rev.* **2024**, *254*, 104814. [[CrossRef](#)]
21. Shi, T.; Chen, Y.; Liu, Y.; Wu, G. Visible and near-infrared reflectance spectroscopy—An alternative for monitoring soil contamination by heavy metals. *J. Hazard. Mater.* **2014**, *265*, 166–176. [[CrossRef](#)] [[PubMed](#)]
22. 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](#)] [[PubMed](#)]
23. Huang, Y.; Chen, W.; Tan, W.; Deng, Y.; Yang, C.; Zhu, X.; Shen, J.; Liu, N. Transfer learning for enhancing the generality of leaf spectroscopic models. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *139*, 104481. [[CrossRef](#)]
24. Padarian, J.; Minasny, B.; McBratney, A.B. Machine learning and soil sciences: A review. *Soil* **2020**, *6*, 35–52. [[CrossRef](#)]
25. Rodriguez-Galiano, V.; Sanchez-Castillo, M.; Chica-Olmo, M.; Chica-Rivas, M. Machine learning predictive models for mineral prospectivity. *Ore Geol. Rev.* **2015**, *71*, 804–818. [[CrossRef](#)]
26. Tan, K.; Chen, L.; Wang, H.; Liu, Z.; Ding, J.; Wang, X. Estimation of heavy metal distribution in soil from airborne hyperspectral imagery. *J. Environ. Manag.* **2023**, *347*, 119196. [[CrossRef](#)]
27. Ma, X.; Wang, J.; Zhou, K.; Zhang, W.; Zhang, Z.; De Maeyer, P.; Van de Voorde, T. Data-driven estimation of metal elements using hyperspectral and geochemical data. *Ore Geol. Rev.* **2024**, *165*, 105877. [[CrossRef](#)]
28. Zhang, T.; Fu, Q.; Tian, R.; Zhang, Y.; Sun, Z. A spectrum contextual self-attention deep learning network for hyperspectral inversion of soil metals. *Ecol. Indic.* **2023**, *152*, 110351. [[CrossRef](#)]
29. Liang, H.; Song, Y.; Dai, Z.; Liu, H.; Zhong, K.; Feng, H.; Xu, L. Soil nitrogen and pH estimation based on GNNW-XGBoost. *Spectrochim. Acta A* **2025**, *330*, 125716. [[CrossRef](#)]
30. Wang, Y.; Zou, B.; Li, S.; Tian, R.; Zhang, B.; Feng, H.; Tang, Y. A hierarchical residual correction-based hyperspectral inversion method. *J. Hazard. Mater.* **2024**, *479*, 135699. [[CrossRef](#)] [[PubMed](#)]
31. Zhao, D.; Wang, J.; Miao, J.; Zhen, J.; Wang, J.; Gao, C.; Jiang, J.; Wu, G. Spectral features of Fe and organic carbon in estimating heavy metals. *Geoderma* **2022**, *426*, 116093. [[CrossRef](#)]
32. Zhang, S.; Chen, Y.; Zhang, Z.; Wang, S.; Wu, Z.; Hong, Y.; Wang, Y.; Hou, H.; Hu, Z.; Fei, T. VNIR estimation of heavy metals concentrations in soil. *Catena* **2022**, *219*, 106585. [[CrossRef](#)]
33. He, F.; Gu, L.; Wang, T.; Zhang, Z. Geo-ecological environmental evaluation using spatiotemporal big data. *J. Clean. Prod.* **2017**, *142*, 854–866. [[CrossRef](#)]
34. Mitran, T.; Gunnam, J.R.S.; Gourigari, S.; Kandrika, S. Assessment of soil heavy metals over an industrial area. *J. Geochem. Explor.* **2024**, *257*, 107379. [[CrossRef](#)]
35. Xiang, Z.; Wu, S.; Zhu, L.; Yang, K.; Lin, D. Pollution characteristics and source apportionment of heavy metals. *J. Environ. Sci.* **2024**, *143*, 23–34. [[CrossRef](#)]
36. Soriano, A.; Pallarés, S.; Pardo, F.; Vicente, A.; Sanfeliu, T.; Bech, J. Deposition of heavy metals in soils of an industrialised area. *J. Geochem. Explor.* **2012**, *113*, 36–44. [[CrossRef](#)]
37. Cai, Y.; Peng, C.; Qiu, S.; Li, Y.; Gao, Y. Determination of soil microbial biomass carbon. *Commun. Soil Sci. Plant Anal.* **2011**, *42*, 2824–2834. [[CrossRef](#)]
38. Liu, Y.; Fan, Y.; Fan, J.; Yue, J.; Chen, R.; Ma, Y.; Bian, M.; Yang, F.; Feng, H. Combining multiple spectral preprocessing and wavelength optimization methods improves potato aboveground biomass estimation. *Inf. Process. Agric.* **2025**, *12*, 511–521. [[CrossRef](#)]
39. Wang, T.; Zhang, Y.; Li, F.; Cao, N. Spectral estimation accuracy of plant phosphorus content. *Comput. Electron. Agric.* **2025**, *234*, 110205. [[CrossRef](#)]
40. Huang, Y.; Jiang, D.; Zhuang, D.; Fu, J. Evaluation of hyperspectral indices for chlorophyll-a estimation. *Int. J. Environ. Res. Public Health* **2010**, *7*, 2437–2451. [[CrossRef](#)] [[PubMed](#)]
41. Shin, H.; Yu, J.; Wang, L.; Jeong, Y.; Kim, J. Spectral interference of heavy metals on moisture signals. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 2266–2275.

42. Geng, J.; Lv, J.; Pei, J.; Liao, C.; Tan, Q.; Wang, T.; Fang, H.; Wang, L. Prediction of soil organic carbon using hyperspectral data. *Comput. Electron. Agric.* **2024**, *220*, 108905. [\[CrossRef\]](#)
43. Greenacre, M.; Groenen, P.; Hastie, T.; d'Enza, A.; Markos, A.; Tuzhilina, E. Principal component analysis. *Nat. Rev. Methods Primers* **2022**, *2*, 100. [\[CrossRef\]](#)
44. Hasan, B.; Abdulazeez, A.M. A review of principal component analysis algorithm. *J. Soft Comput. Data Min.* **2021**, *2*, 20–30.
45. Bhagat, S.; Tung, T.; Yaseen, Z. Heavy metal contamination prediction using ensemble model. *J. Hazard. Mater.* **2021**, *403*, 123492. [\[CrossRef\]](#) [\[PubMed\]](#)
46. Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K.; Mitchell, R.; Cano, I.; Zhou, T.; et al. Extreme Gradient Boosting [R package xgboost version 1.2. 0.1]. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual, 6–10 July 2020; pp. 13–17.
47. Grekousis, G. Geographical-XGBoost: A new ensemble model. *J. Geogr. Syst.* **2025**, *27*, 169–195. [\[CrossRef\]](#)
48. Xu, X.; Wang, X.; Yang, P.; Meng, Y.; Yu, D.; Li, C. Mapping soil salt contents using satellite imagery. *Catena* **2023**, *223*, 106900. [\[CrossRef\]](#)
49. Xia, K.; Wu, T.; Zhang, S.; Wang, S.; Li, X.; Shen, Q.; Li, T. High-precision estimation of soil organic matter using spectroscopy. *Geoderma* **2024**, *445*, 116877. [\[CrossRef\]](#)
50. Jia, Y.; Jin, S.; Savi, P.; Gao, Y.; Tang, J.; Chen, Y.; Li, W. GNSS-R soil moisture retrieval using XGBoost. *Remote Sens.* **2019**, *11*, 1655. [\[CrossRef\]](#)
51. Hengl, T.; Leenaars, J.; Shepherd, K.; Walsh, M.; Heuvelink, G.; Mamo, T.; Tilahun, H.; Berkhout, E.; Cooper, M.; Fegraus, E. Soil nutrient maps of Sub-Saharan Africa: Assessment of soil nutrient content at 250 m spatial resolution using machine learning. *Nutr. Cycl. Agroecosyst.* **2017**, *109*, 77–102. [\[CrossRef\]](#) [\[PubMed\]](#)
52. Bellon-Maurel, V.; Fernandez-Ahumada, E.; Palagos, B.; Roger, J.-M.; McBratney, A. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. *TrAC Trends Anal. Chem.* **2010**, *29*, 1073–1081. [\[CrossRef\]](#)
53. Xu, X.; Wang, Z.; Song, X.; Zhan, W.; Yang, S. A remote sensing-based strategy for mapping potentially toxic elements of soils: Temporal-spatial-spectral covariates combined with random forest. *Environ. Res.* **2024**, *240*, 117570. [\[CrossRef\]](#) [\[PubMed\]](#)
54. Kurkcuoglu, M. Analysis of the energy justice in natural gas distribution with multiscale geographically weighted regression (MGWR). *Energy Rep.* **2023**, *9*, 325–337. [\[CrossRef\]](#)
55. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
56. Ye, M.; Zhu, L.; Li, X.; Ke, Y.; Huang, Y.; Chen, B.; Yu, H.; Li, H.; Feng, H. Estimation of soil arsenic concentration using a geographically weighted XGBoost model based on hyperspectral data. *Sci. Total Environ.* **2023**, *858*, 159798. [\[CrossRef\]](#) [\[PubMed\]](#)
57. Xu, X.; Chen, Y.; Wang, M.; Wang, S.; Li, K.; Li, Y. Improving estimates of soil salt content by using two-date image spectral changes in Yinbei, China. *Remote Sens.* **2021**, *13*, 4165. [\[CrossRef\]](#)
58. Wu, D.; Liu, H.; Wu, J.; Gao, X.; Nyasha, N.K.; Cai, G.; Zhang, W. Bi-directional pollution characteristics and ecological health risk assessment of heavy metals in soil and crops. *Int. J. Environ. Res. Public Health* **2022**, *19*, 9669. [\[CrossRef\]](#) [\[PubMed\]](#)
59. GB 15618–2018; Soil Environmental Quality Risk Control Standard for Soil Contamination of Agricultural Land. Ministry of Ecology and Environment of the People's Republic of China: Beijing, China, 2018.
60. Rossel, R.V.; Behrens, T. Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma* **2010**, *158*, 46–54. [\[CrossRef\]](#)
61. Ben-Dor, E.; Inbar, Y.; Chen, Y. The reflectance spectra of organic matter in the visible, near-infrared and shortwave infrared region (400–2500 nm) during a controlled decomposition process. *Remote Sens. Environ.* **1997**, *61*, 1–15. [\[CrossRef\]](#)
62. Xu, M.; Wang, S.; Zhang, S.; Li, F.; Ma, C.; Cheng, Z. Hyperspectral reflectance models for retrieving heavy metal content: Application in archaeological soil. *J. Infrared Millim. Waves* **2011**, *30*, 109–114. [\[CrossRef\]](#)
63. Sun, W.; Zhang, X.; Zou, B.; Wu, T. Exploring the potential of spectral classification in estimation of soil contaminant elements. *Remote Sens.* **2017**, *9*, 632. [\[CrossRef\]](#)
64. Seifi, M.; Ahmadi, A.; Neyshabouri, M.R.; Taghizadeh-Mehrjardi, R.; Bahrami, H.-A. Remote and Vis-NIR spectroscopy for soil salinization estimation. *Remote Sens. Appl. Soc. Environ.* **2020**, *20*, 100398. [\[CrossRef\]](#)
65. Wang, J.; Cui, L.; Gao, W.; Shi, T.; Chen, Y.; Gao, Y. Prediction of low heavy metal concentrations in agricultural soils using visible and near-infrared reflectance spectroscopy. *Geoderma* **2014**, *216*, 1–9. [\[CrossRef\]](#)
66. Xie, S.; Ding, F.; Chen, S.; Wang, X.; Li, Y.; Ma, K. Prediction of soil organic matter content based on characteristic band selection method. *Spectrochim. Acta A* **2022**, *273*, 120949. [\[CrossRef\]](#)
67. Jiang, Q.; Liu, M.; Wang, J.; Liu, F. Feasibility of using visible and near-infrared reflectance spectroscopy to monitor heavy metal contaminants in urban lake sediment. *Catena* **2018**, *162*, 72–79. [\[CrossRef\]](#)
68. Chen, H.; Jilili, A.; Liu, W.; Chen, J. Correlation between heavy metals, organic matter and pH value in soils along the Bortala River. *Res. Soil Water Conserv.* **2016**, *23*, 210–213.

69. Moosavi, V.; Zuravand, G.; Shamsi, S. Cluster-based local modeling meets deep learning: A novel approach to soil moisture estimation. *J. Hydrol.* **2024**, *635*, 131161. [\[CrossRef\]](#)
70. Shi, L.; Daly, K.; O'Rourke, S. Estimating mineral-associated organic carbon saturation using MIR spectroscopy. *Geoderma* **2025**, *454*, 117181. [\[CrossRef\]](#)
71. Ramirez-Lopez, L.; Schmidt, K.; Behrens, T.; Van Wesemael, B.; Dematté, J.A.; Scholten, T. Sampling optimal calibration sets in soil infrared spectroscopy. *Geoderma* **2014**, *226*, 140–150. [\[CrossRef\]](#)
72. Nawar, S.; Mouazen, A. Optimal sample selection for measurement of soil organic carbon using on-line vis-NIR spectroscopy. *Comput. Electron. Agric.* **2018**, *151*, 469–477. [\[CrossRef\]](#)
73. Fu, Y.; Li, F.; Guo, S.; Zhao, M. Cadmium concentration and fluxes in agricultural soil downstream of sewage irrigation area. *J. Hazard. Mater.* **2021**, *412*, 125203. [\[CrossRef\]](#) [\[PubMed\]](#)
74. Tóth, G.; Hermann, T.; Szatmári, G.; Pásztor, L. Maps of heavy metals in soils of the European Union and priority areas for assessment. *Sci. Total Environ.* **2016**, *565*, 1054–1062. [\[CrossRef\]](#) [\[PubMed\]](#)
75. Tan, K.; Ma, W.; Chen, L.; Wang, H.; Du, Q.; Du, P.; Yan, B.; Liu, R.; Li, H. Estimating distribution trends of soil heavy metals in mining areas using hyperspectral imagery. *J. Hazard. Mater.* **2021**, *401*, 123288. [\[CrossRef\]](#) [\[PubMed\]](#)
76. Li, F.; Xu, L.; Jin, C.; Wang, H. Structure of multi-stage composite genetic algorithm (MSC-GA) and its performance. *Expert Syst. Appl.* **2011**, *38*, 8929–8937. [\[CrossRef\]](#)
77. Shi, T.; Chen, Y.; Liu, H.; Wang, J.; Wu, G. Soil organic carbon estimation using visible–near-infrared spectroscopy with feature selection. *Appl. Spectrosc.* **2014**, *68*, 831–837. [\[CrossRef\]](#) [\[PubMed\]](#)
78. Xu, X.; Wang, Y.; Dai, X.; Shen, Q.; Wu, Q.; Wang, Z.; Cao, J. Improving satellite-based estimation and mapping of soil lead using enhanced spectral features and XGBoost model. *Remote Sens.* **2026**, *18*, 1446. [\[CrossRef\]](#)
79. Li, J.; Zheng, K.; Gao, L.; Ni, L.; Huang, M.; Chanussot, J. Model-Informed Multistage Unsupervised Network for Hyperspectral Image Super-Resolution. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5516117. [\[CrossRef\]](#)
80. Wittstruck, L.; Waske, B.; Jarmer, T. Multi-Modal Vision Transformer for high-resolution soil texture prediction of German agricultural soils using remote sensing imagery. *Remote Sens. Environ.* **2025**, *331*, 114985. [\[CrossRef\]](#)
81. Helfenstein, A.; Mulder, V.; Heuvelink, G.B.; Hack-ten Broeke, M.J. Three-dimensional space and time mapping reveals soil organic matter decreases across anthropogenic landscapes in the Netherlands. *Commun. Earth Environ.* **2024**, *5*, 130. [\[CrossRef\]](#)

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.