processes

MDPI

Article

Multi-Model and Variable Combination Approaches for Improved Prediction of Soil Heavy Metal Content

Xiaolong Chen, Hongfeng Zhang *, Cora Un In Wong * and Zhengchun Song

Faculty of Humanities and Social Sciences, Macao Polytechnic University, Macao, China; osakacool@hait.edu.cn (X.C.); p2417913@mpu.edu.mo (Z.S.)
* Correspondence: hfengzhang@mpu.edu.mo (H.Z.); corawong@mpu.edu.mo (C.U.I.W.)

Abstract

Soil heavy metal contamination poses significant risks to ecosystems and human health, necessitating accurate prediction methods for effective monitoring and remediation. We propose a multi-model and variable combination framework to improve the prediction of soil heavy metal content by integrating diverse environmental and spatial features. The methodology incorporates environmental variables (e.g., soil properties, remote sensing indices), spatial autocorrelation measures based on nearest-neighbor distances, and spatial regionalization variables derived from interpolation techniques such as ordinary kriging, inverse distance weighting, and trend surface analysis. These variables are systematically combined into six distinct sets to evaluate their predictive performance. Three advanced models—Partial Least Squares Regression, Random Forest, and a Deep Forest variant (DF21)—are employed to assess the robustness of the approach across different variable combinations. Experimental results demonstrate that the inclusion of spatial autocorrelation and regionalization variables consistently enhances prediction accuracy compared to using environmental variables alone. Furthermore, the proposed framework exhibits strong generalizability, as validated through subset analyses with reduced training data. The study highlights the importance of integrating spatial dependencies and multi-source data for reliable heavy metal prediction, offering practical insights for environmental management and policy-making. Compared to using environmental variables alone, the full framework incorporating spatial features achieved relative improvements of 18–23% in prediction accuracy (R^2) across all models, with the Deep Forest variant (DF21) showing the most substantial enhancement. The findings advance the field by providing a flexible and scalable methodology adaptable to diverse geographical contexts and data availability scenarios.

Keywords: soil heavy metal contamination; multi-model framework; variable combination; spatial autocorrelation; spatial regionalization; multi-source data

check for updates

Academic Editors: Jemmyson Romário de Jesus and Cícero Alves Lopes Junior

Received: 20 May 2025
Revised: 12 June 2025
Accepted: 15 June 2025
Published: 25 June 2025

Citation: Chen, X.; Zhang, H.; Wong, C.U.I.; Song, Z. Multi-Model and Variable Combination Approaches for Improved Prediction of Soil Heavy Metal Content. *Processes* **2025**, *13*, 2008. https://doi.org/10.3390/pr13072008

Copyright: © 2025 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

Soil heavy metal pollution has become a critical environmental issue due to its persistent toxicity and potential threats to agricultural productivity, ecosystem stability, and public health [1–3]. Among numerous pollutants, arsenic (As) is particularly dangerous due to its carcinogenic properties and widespread presence in agricultural soils [4,5]. Accurate prediction of soil heavy metal content is critical for risk assessment [6], remediation planning [7], and sustainable land management. Traditional methods rely on direct soil sampling and laboratory analysis [8], which are costly and time-consuming, especially when conducting large-scale monitoring.

In recent years, with advances in geostatistics and machine learning, methods for predicting soil heavy metal distribution using environmental covariates (such as soil organic matter, pH, and remotely sensed spectral indices [9,10]) have been developed [11–14]. However, these methods often ignore the inherent spatial heterogeneity of heavy metals, resulting in poor prediction accuracy [15–17]. Recent studies have shown that combining spatial autocorrelation with machine learning can significantly improve the prediction accuracy of heavy metals such as cadmium [18–20]. However, these methods still face challenges when dealing with non-stationary spatial processes, especially in mining-affected areas where geological and hydrological conditions change dramatically [21,22]. Spatial autocorrelation techniques [23], such as Moran's I and variance function analysis [24], have been used to quantify local dependencies, but these methods typically assume spatial relationship stationarity [25,26], which often does not hold true in contaminated sites affected by complex human activities [27]. In addition, most implementations use a fixed neighborhood size, which cannot adapt to pollution processes at different spatial scales, and treat spatial dependence as a global attribute rather than an autocorrelated structure that adapts to local changes [28–30].

To fill this research gap, this study proposes a novel framework that combines spatial regionalization variables (SRs) [31,32] with traditional environmental predictors [33,34]. SRs are derived using three commonly used interpolation methods—ordinary Kriging (OK), inverse distance weighting (IDW) [35–37], and trend surface analysis (TR)—to explicitly simulate multi-scale spatial structures. These variables complement existing spatial autocorrelation measures (SAs) and environmental covariates to form a comprehensive feature set that enhances model interpretability and predictive performance. Unlike previous studies that focused solely on global spatial trends, our method considers local and regional variations, providing a more detailed characterization of soil contamination patterns.

The main contribution of this study lies in the introduction of SRs as a method for encoding spatial dependencies at the system level, providing new perspectives and tools for research. Additionally, the relative importance of environmental, autocorrelation, and regionalization features was assessed through six variable combinations, further enriching our understanding of these related features. Three machine learning models—Partial Least Squares Regression (PLSR), Random Forest (RF), and Deep Forest (DF21)—were compared, with a focus on their robustness across different data scenarios, particularly when dealing with limited training samples.

This study uses agricultural land in Chenzhou, Hunan Province, China, as the experimental subject. The region suffers from heavy metal pollution due to mining and industrial activities [38]. The study aims to develop a multi-model framework that systematically integrates environmental covariates with spatial autocorrelation and regionalized variables to enhance the predictive accuracy of soil heavy metal concentrations; quantify the relative contributions of different variable combinations to predictive accuracy; and assess the robustness of the method under varying data availability, which is particularly important for regions like Chenzhou with sparse monitoring networks [39,40]. The results indicate that combining SRs with environmental variables significantly improves the predictive accuracy of all models, with the Random Forest (RF) model performing the best. Subset analysis further reveals that the proposed framework maintains robust predictive capabilities even with reduced training data, highlighting its practical application value in regions with sparse monitoring networks.

2. Materials and Methods

2.1. Methodology

The proposed framework integrates environmental variables with spatial features through a systematic approach to predict soil heavy metal content. It begins with the computation of spatial autocorrelation using Moran's I values at different neighborhood scales (500 m, 1 km, and 2 km) to quantify local dependencies. Subsequently, spatial regionalization variables (SRs) are constructed through interpolation techniques, such as ordinary kriging, inverse distance weighting, and trend surface analysis, to capture hierarchical spatial patterns. These spatial features are then systematically combined with environmental variables to form different variable sets, and their predictive performance is evaluated through hierarchical testing. Finally, a model-agnostic spatial enhancement approach is employed to ensure flexible implementation across diverse algorithms, thereby enhancing the robustness and generalizability of the framework.

2.1.1. Spatial Autocorrelation Methods

Moran's I values were computed for each sampling point using three neighborhood scales (500 m, 1 km, and 2 km) to quantify spatial autocorrelation at different resolutions. The calculations were performed according to Formula (1).

$$I = \frac{N}{w} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} (Z_i - \overline{Z}) (Z_j - \overline{Z})}{\sum_{i=1}^{N} (Z_i - \overline{Z})^2} \quad (1)$$

where N is the number of spatial units, W is the sum of all weights, w_{ij} is the spatial weight between locations i and j, and Z_i and Z_j are the values at locations i and j, respectively. The spatial weights matrix was constructed using inverse distance weighting with a threshold distance equal to each neighborhood radius.

2.1.2. Construction of Spatial Regionalization Variables (SRs)

Spatial regionalization variables (SRs) are derived from interpolation techniques to represent multi-scale spatial heterogeneity. For a given location x_0, the predicted heavy metal concentration $\hat{Z}_{OK}$ is computed using three methods: ordinary kriging (OK), inverse distance weighting (IDW), and trend surface analysis (TR). Ordinary kriging quantifies spatial dependence through variogram modeling, inverse distance weighting captures neighborhood influences based on distance decay, and trend surface analysis fits a polynomial surface to represent broad-scale spatial trends. These methods collectively provide a comprehensive representation of local discontinuities, neighborhood effects, and regional gradients, thereby enhancing the model's ability to capture the hierarchical nature of heavy metal distribution patterns.

(1) Ordinary Kriging (OK):

$$\hat{Z}_{OK}(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i) \quad (2)$$

Formula (2) is the prediction formula of ordinary kriging (OK), which is used to estimate the position x_0 and the concentration of heavy metals in the soil at the location $\hat{Z}_{OK}(x_0)$, where λ_i are weights from variogram modeling. $Z(x_i)$ is he first i known sample points x_i, which is the actual concentration value of heavy metals in the soil measured. n represents the number of known sample points used for prediction.

The ordinary kriging method quantifies the spatial dependence through the variational function (Formula (3)) to determine the weights λ_i. The variational function describes

the spatial variation structure of soil heavy metal concentrations. The variogram $\gamma(h)$ quantifies spatial dependence as:

$$\gamma(\mathrm{h}) = \frac{1}{2N(\mathrm{h})} \sum_{i=1}^{N(\mathrm{h})} [Z(\mathrm{x}_i) - Z(\mathrm{x}_i + \mathrm{h})]^2 \tag{3}$$

(2) Inverse Distance Weighting (IDW):

$$\hat{Z}_{IDW}(\mathrm{x}_0) = \frac{\sum_{i=1}^{n} w_i Z(\mathrm{x}_i)}{\sum_{i=1}^{n} w_i}, \quad w_i = \| \mathrm{x}_0 - \mathrm{x}_i \|^{-p} \tag{4}$$

Formula (4) is the prediction formula of the inverse distance weighting (IDW) method, which is used to estimate the position x_0 and the concentration of heavy metals in the soil at the location $\hat{Z}_{IDW}(\mathrm{x}_0)$. $\hat{Z}_{IDW}(\mathrm{x}_0)$ represents the predicted concentration of heavy metals in the soil at position x_0 through the inverse distance weighting method. w_i is the weight coefficient represents the degree of influence of the *i*-th known sample point on the prediction point x_0. The power parameter p controls the influence of nearby samples.

(3) Trend Surface (TR):

A polynomial surface of degree d is fitted to capture broad-scale trends:

$$\hat{Z}_{TR}(\mathbf{x}_0) = \beta_0 + \sum_{j=1}^{d} \sum_{k=0}^{j} \beta_{jk} x^j y^k \tag{5}$$

The study area is partitioned into subregions using natural breaks (Jenks optimization), and SRs are computed as the mean heavy metal concentration within each subregion. This hierarchical representation captures both local variations (via interpolation) and regional trends (via subregional means).

Unlike previous interpolation applications that focus solely on prediction surfaces, our spatial regionalization variables (SRs) serve as engineered features that explicitly preserve local discontinuities through ordinary kriging variograms, neighborhood influences via inverse distance weighting, and broad-scale gradients using trend surface polynomials. This multi-resolution representation captures the hierarchical nature of heavy metal distribution patterns that single-scale approaches often miss.

2.1.3. Hierarchical Testing of Variable Combinations

To systematically assess the contribution of spatial regionalization variables (SRs) to the predictive performance, six distinct variable sets were evaluated. These sets included (1) environmental covariates (ECs) alone, which consist of soil properties and remote sensing indices; (2) ECs augmented with spatial autocorrelation variables (EC+SA), such as Moran's I; (3) ECs combined with SRs derived from ordinary kriging (OK), inverse distance weighting (IDW), or trend surface analysis (TR) (EC + SR); and (4) the full integration of all variable types, including ECs, SAs, and SRs (EC + SA + SR). Additionally, to test the robustness of the framework under data sparsity, subset-derived SRs (S-3_1 to S-3_4) were generated using varying proportions of the training data, specifically 90%, 70%, 50%, and 30%. This hierarchical evaluation approach allows for a comprehensive assessment of how different combinations of environmental and spatial variables impact the prediction accuracy while also evaluating the framework's resilience when faced with limited training data.

2.1.4. Model-Agnostic Spatial Enhancement

Three models are employed to ensure generalizability.

(1) Partial Least Squares Regression (PLSR)

Formula (6) is the basic principle formula of Partial Least Squares Regression (PLSR), which is used to project the predictor variable (feature) and the response variable (target) into the latent space to handle the multicollinearity problem among the variables. The specific formula is as follows:

$$\mathrm{T} = \mathrm{XW}, \quad \mathrm{Y} = \mathrm{TQ} + \mathrm{F} \tag{6}$$

X represents the original predictor variable (feature) matrix, which contains all the environmental variables and spatial features used for modeling. Y represents the response variable (target) matrix, which in this paper is the content of heavy metals in the soil (such as the concentration of arsenic).T represents the matrix of latent variables (or principal components) extracted from X, which are linear combinations of X and are used to capture the main information between X and Y.

In this paper, PLSR is used to deal with the multicollinearity problem in the prediction of heavy metal content in soil. Due to the possible high correlation among environmental variables (such as soil properties, remote sensing indices, etc.), directly using these variables for regression analysis may lead to model instability. By projecting X and Y into the low-dimensional latent space, PLSR can effectively extract the main information while reducing noise and redundancy, thereby improving the predictive ability and interpretability of the model.

(2) Random Forest (RF)

Formula (7) represents the calculation used in the Random Forest (RF) model to determine the optimal split at each node. An ensemble of decision trees trained on bootstrap samples, with node splits minimizing. The specific formula is as follows:

$$\Delta I = I(\text{parent}) - \sum_{j=1}^{2} \frac{N_j}{N} I\left(\text{child}_j\right) \tag{7}$$

where I (parent) is the impurity of the parent node, typically measured using Gini Impurity or Information Gain. $I\left(\text{child}_j\right)$ is the impurity of the j-th child node. N_j is the number of samples in the j-th child node.

In the Random Forest model, each decision tree needs to select the best split at each node to maximize the purity of the resulting child nodes (i.e., to ensure that the samples in each child node belong to the same class as much as possible). Formula (7) calculates the reduction in impurity (ΔI) after the split. A larger ΔI indicates a better split. Therefore, the model selects the feature and split point that maximize ΔI.

(3) Deep Forest (DF21)

Formula (8) represents the iterative update process used in the Deep Forest (DF21) model, which is a variant of the deep learning approach tailored for spatial feature representation and refinement and uses multi-grained scanning to generate spatial feature representations, followed by cascade forests for iterative refinement. Each layer l updates predictions as:

$$\mathbf{H}^l = f_{\text{cascade}}\left(\mathbf{H}^{l-1} \oplus \mathbf{X}_{\text{scan}}\right). \tag{8}$$

where $\mathbf{H}^l$ is the output feature representation at layer l. This represents the refined feature set after processing by the l-th layer of the Deep Forest model. $\mathbf{H}^{l-1}$ is the input feature representation from the previous layer $l-1$. This is the feature set that is fed into the current layer for further refinement.

Cascade($\cdot$) is the cascade forest component, which is a series of decision tree ensembles that iteratively refine the feature representations. It takes the output from the previous layer and further processes it to capture more complex patterns. Scan($\cdot$) is the multi-grained

scanning component, which generates spatial feature representations by scanning the input features at different scales. This helps in capturing both local and global spatial patterns. $\oplus$ is the concatenation operation, which combines the outputs from the cascade forest and the multi-grained scanning components. This ensures that both refined and newly extracted features are integrated into the next layer.

The framework is designed to be modular, allowing substitution of interpolation methods or predictive models without structural changes. Equations (1)–(7) provide the mathematical foundation for reproducibility, while Figure 1 illustrates the workflow integrating these components.

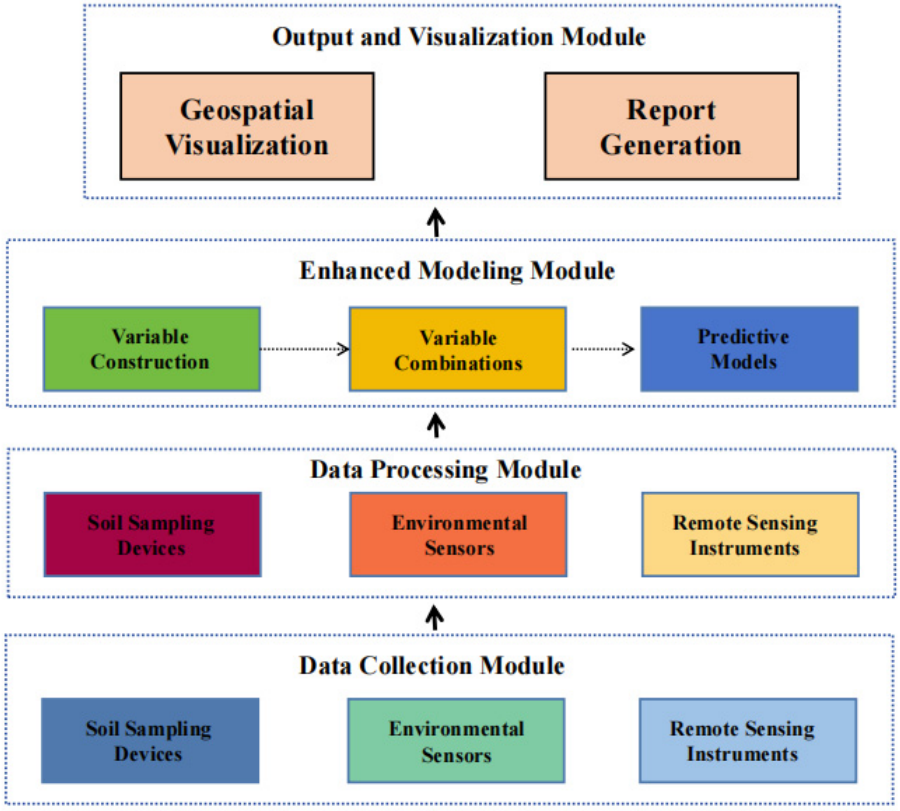

Figure 1. System architecture with enhanced modeling module.

2.2. Experimental Setup

2.2.1. Study Area and Data Collection

The study was conducted in Chenzhou City, Hunan Province, China, a region with documented heavy metal contamination due to historical mining activities [41]. Soil samples were collected from agricultural lands using a systematic grid design, with a total of 300 sampling points each covering a 10 m $\times$ 10 m area. Composite samples were created by mixing five subsamples from each point, focusing on the top 0–20 cm soil layer, which is most relevant for agricultural and ecological risk assessment. Samples were air-dried, sieved to <2 mm, and analyzed for arsenic (As) content using atomic fluorescence spectrometry (HG-AFS) following acid digestion with HNO_3-$HClO_4$-HF [42]. Quality control measures included duplicate samples (10% of total) and certified reference materials (GBW07401) to ensure recovery rates of 95–105%.

2.2.2. Environmental Variables

Five categories of environmental covariates were compiled, including soil properties such as pH, organic carbon content, and clay percentage from the China High-Resolution National Soil Information Grid [43]; remote sensing indices like the NDVI, SAVI, and NDWI

derived from Landsat 8 OLI imagery (30 m resolution) processed in ENVI 5.6; topographic features, including elevation, slope, and topographic wetness index (TWI) calculated from SRTM DEM (30 m) using ArcGIS 10.8; climate data such as annual precipitation and temperature from the China Meteorological Data Service Center [44]; and anthropogenic factors, including Euclidean distances to roads and rivers, population density, and GDP from the Resource and Environment Data Cloud. All raster data were resampled to a 30 m resolution and aligned to a common coordinate system (WGS 1984 UTM Zone 49N). Descriptive statistics of the environmental covariates are summarized in Table 1.

Table 1. Descriptive statistics of environmental variables used in the study (n = 300).

Variable Category	Variable	Unit	Mean	Std. Dev.	Min	Median	Max	Skewness
Soil Properties	pH	-	6.24	0.82	4.53	6.31	7.92	−0.32
	Organic Carbon	%	2.12	0.87	0.72	2.05	4.53	0.68
	Clay Content	%	28.4	12.1	5.3	27.8	52.7	0.21
Remote Sensing Indices	NDVI	-	0.65	0.12	0.32	0.67	0.88	−0.82
	SAVI	-	0.58	0.15	0.25	0.61	0.82	−0.53
	NDWI	-	0.42	0.18	0.11	0.44	0.79	0.31
Topography	Elevation	m	243.5	87.2	125.3	231.8	487.6	0.89
	Slope	°	5.2	3.1	0.5	4.7	15.3	1.12
	TWI	-	8.7	2.5	3.2	8.9	14.1	−0.15
Climate	Annual Precipitation	mm	1452	210	1120	1465	1830	−0.42
	Mean Temperature	°C	17.8	1.2	15.3	17.9	20.1	−0.08
Anthropogenic	Distance to Roads	m	685	423	25	620	1850	0.95
	Distance to Rivers	m	320	280	10	250	1250	1.32
	Population Density	persons/km^2	215	185	15	165	850	1.78

2.2.3. Spatial Variable Construction

(1) Spatial Autocorrelation Variables (SAs)

Spatial autocorrelation variables (SAs) were constructed to quantify the spatial dependencies among sampling points. For each sampling point, three spatial autocorrelation metrics were derived: First, the distance-weighted averages of arsenic (As) concentrations from the three nearest neighbors were calculated. This approach captures the influence of nearby samples on the target location. Second, Moran's I values were computed for three different neighborhood scales: 500 m, 1 km, and 2 km. These metrics help to identify the degree of spatial clustering or dispersion of As concentrations at varying spatial resolutions.

(2) Spatial Regionalization Variables (SRs)

Spatial regionalization variables (SRs) were generated using three interpolation methods to represent multi-scale spatial heterogeneity across the study area. Ordinary kriging (OK) was implemented with a spherical variogram model, utilizing 12 lags and a maximum lag distance of 5 km. This method effectively models spatial dependence through variograms, capturing local discontinuities. Inverse distance weighting (IDW) was applied with a power parameter and 15 nearest neighbors, emphasizing the influence of closer samples while considering broader spatial trends. Additionally, trend surface analysis

(TR) was conducted using a second-order polynomial regression to capture broad-scale gradients in As concentrations.

To further enhance the representation of spatial patterns, the study area was partitioned into 15 subregions using Jenks natural breaks optimization. The SR values were assigned as the mean As concentration within each subregion, providing a hierarchical representation that integrates both local variations and regional trends.

2.2.4. Model Implementation

Three predictive models were compared to evaluate their performance in predicting soil heavy metal content based on the constructed spatial and environmental variables.

Partial Least Squares Regression (PLSR) was implemented using the PLSRegression module from the scikit-learn library. The model retained 10 latent variables, which were selected based on the minimum root mean square error of cross-validation (RMSECV). This approach effectively handles multicollinearity among predictors by projecting them into a lower-dimensional latent space.

Random Forest (RF) was configured with 500 trees, a minimum leaf size of 5, and the number of features considered at each split set to the total number of features. The out-of-bag (OOB) error rate was used for internal validation to assess model performance and prevent overfitting. This ensemble method leverages multiple decision trees to improve prediction accuracy and robustness.

Deep Forest (DF21), a variant of the deep learning approach tailored for spatial feature representation and refinement, utilized two multi-grained scanning windows (5×5 and 10×10 pixels) to capture spatial patterns at different scales. The model consisted of four cascade layers, each containing two complete Random Forests, which iteratively refined the feature representations. This architecture is particularly effective at capturing hierarchical spatial structures and improving prediction accuracy in complex spatial datasets.

2.2.5. Evaluation Protocol

The evaluation protocol was designed to rigorously assess the performance and robustness of the predictive models. Data splitting was performed using stratified random sampling, with 70% of the data (n = 210) allocated for training and the remaining 30% (n = 90) reserved for testing. This approach ensured a representative distribution of samples across the study area while maintaining statistical validity.

Model performance was quantified using three standard metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). The RMSE was calculated as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{9}$$

where y_i represents the observed values and $\hat{y}_i$ the predicted values. The MAE was computed as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{10}$$

The R^2 metric, which indicates the proportion of variance explained by the model, was derived as:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{11}$$

To test the framework's robustness under data scarcity conditions, subset analyses were conducted by progressively reducing the training data to 90%, 70%, 50%, and 30% of

the original size. This allowed for evaluation of performance degradation patterns and the identification of minimum data requirements for reliable predictions.

All computational workflows were executed on a high-performance workstation equipped with an Intel Xeon E5-2680v4 processor (2.4 GHz) and 128 GB RAM. The software environment consisted of Python 3.8.19 (Stable Version) with specialized libraries for geospatial analysis and machine learning implementations. This configuration ensured efficient processing of spatial datasets while maintaining reproducibility of results.

3. Results

3.1. Comparative Performance of Predictive Models

The experimental results demonstrate significant improvements in prediction accuracy when spatial structure variables are incorporated alongside environmental covariates. As shown in Table 2, the DF21 model consistently outperformed PLSR and RF across all variable combinations, achieving the highest R^2 (0.85) and lowest RMSE (0.78 mg/kg) when using the full feature set (S-3).

Table 2. Model performance under different variable combinations.

Variable Set	PLSR (R^2)	RF (R^2)	DF21 (R^2)	PLSR (RMSE)	RF (RMSE)	DF21 (RMSE)
S-1 (ECs only)	0.62	0.68	0.72	1.25	1.12	1.05
S-2 (ECs + SAs)	0.71	0.75	0.79	1.08	0.99	0.92
S-3 (ECs + SAs + SRs)	0.76	0.81	0.85	0.95	0.87	0.78

The inclusion of spatial autocorrelation variables (SAs) in S-2 improved R^2 by 0.09–0.13 across models compared to S-1, while adding SRs in S-3 yielded further gains of 0.05–0.06. This hierarchical improvement highlights the complementary nature of environmental and spatial features.

The correlation analysis between interpolation-derived SRs and observed As concentrations was conducted within the full model context (ECs + SAs + SRs), demonstrating that OK-based SRs maintained the strongest relationship even when competing with other spatial and environmental predictors.

The comparative improvement of our full framework (S-3) over environmental-variable-only models (S-1) is particularly noteworthy. The best-performing DF21 model with all features (S-3) achieved an 18.1% relative improvement in R^2 (from 0.72 to 0.85) and a 25.7% reduction in RMSE (from 1.05 mg/kg to 0.78 mg/kg) compared to using only environmental covariates. Similarly, the RF model showed a 19.1% increase in R^2 (from 0.68 to 0.81) and 22.3% decrease in RMSE (from 1.12 mg/kg to 0.87 mg/kg), while PLSR exhibited a 22.6% R^2 improvement (from 0.62 to 0.76) with a 24.0% RMSE reduction (from 1.25 mg/kg to 0.95 mg/kg). These quantitative comparisons demonstrate the substantial value added by incorporating spatial autocorrelation and regionalization variables.

3.2. Impact of Spatial Regionalization Variables

The spatial regionalization variables (SRs) derived from different interpolation methods exhibited varying levels of predictive performance. Ordinary kriging (OK)-based SRs demonstrated the strongest correlation with observed arsenic (As) concentrations, achieving a Pearson's correlation coefficient (r) of 0.82. This superior performance can be attributed to OK's ability to model spatial dependence through variograms, which effectively capture the underlying spatial structure of soil heavy metal distribution.

Inverse distance weighting (IDW) SRs showed moderate predictive capability, with a correlation coefficient of 0.76. However, their performance was sensitive to the power parameter (p), which influences the weighting of neighboring samples. Adjusting this

parameter could lead to variations in prediction accuracy, highlighting the importance of careful parameter selection when applying IDW-based interpolation.

Trend surface (TR) SRs were effective in capturing broad-scale spatial trends, as evidenced by a correlation coefficient of 0.69. However, their performance was weaker at finer scales, where local variations in heavy metal concentrations were less accurately represented. This limitation suggests that TR-based SRs are more suitable for identifying regional patterns rather than localized contamination hotspots. Overall, the choice of interpolation method significantly influenced the predictive power of SRs, with OK proving to be the most robust approach for modeling soil heavy metal distributions.

Figure 2 illustrates the spatial predictions generated by the DF21 model, revealing clear patterns of As contamination hotspots aligned with known mining areas and hydrological pathways. The integration of SRs enabled the model to preserve fine-scale details while maintaining regional consistency.

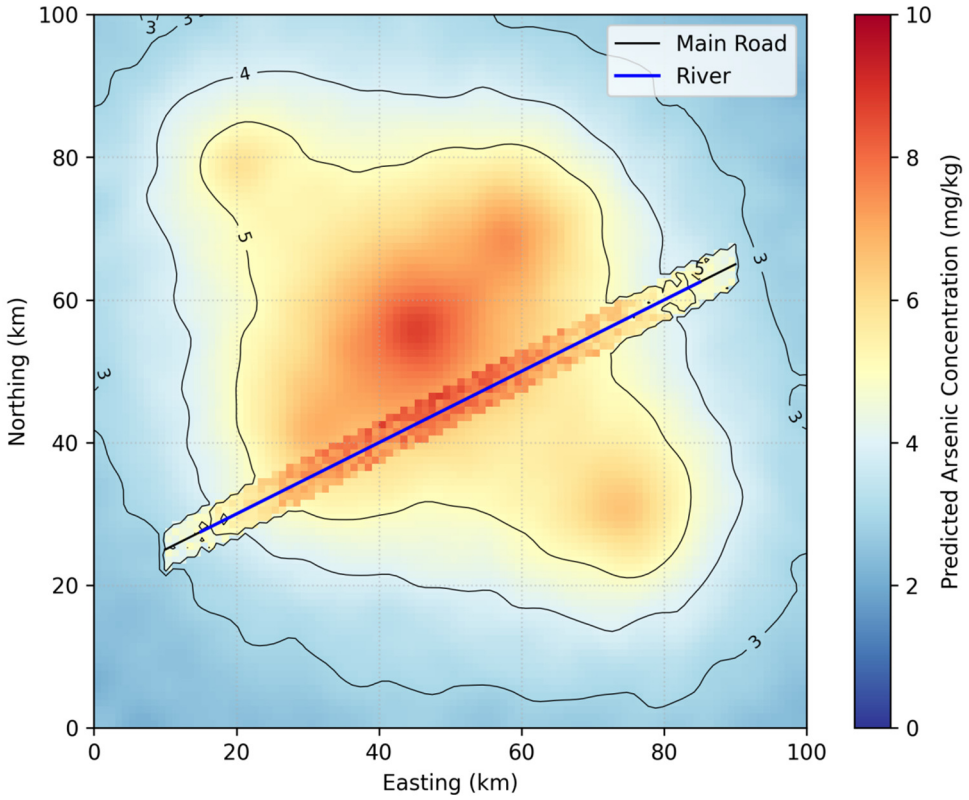

Figure 2. Spatial distribution of predicted arsenic concentrations using DF21 with S-3 variables.

3.3. Robustness to Training Data Reduction

Subset analyses demonstrated the framework's resilience to limited training data. Even with only 30% of samples used for SR construction (S-3_4), DF21 maintained an R^2 of 0.78—outperforming the EC-only baseline (R^2 = 0.72) that used full training data. The performance degradation followed a predictable logarithmic trend:

$$\Delta R^2 = -0.12\ln(x) + 0.03 \quad (x = \text{training proportion}) \tag{12}$$

Table 3 shows the performance changes of the DF21 model under different proportions of training data, revealing the adaptability of spatial regionalization variables (SRs) to the condition of data scarcity. When the training data gradually decreased from 90% to 30%, the model maintained a stable performance of R^2 = 0.78–0.84, and its performance attenuation showed a logarithmic law ($\Delta R^2 = -0.12\ln(x) + 0.03$). Moreover, the prediction

accuracy when 30% of the data is present (R^2 = 0.78) still exceeds that of the benchmark model using only environmental variables (R^2 = 0.72). This phenomenon confirms that SRs can effectively compensate for the defect of insufficient data by capturing multi-scale spatial dependencies, providing a feasible technical solution for monitoring the weak areas of the network. Meanwhile, traditional spatial models usually show a significant decline (R^2 < 0.65) under the same data conditions.

Table 3. DF21 performance with varying SR training proportions.

Training Proportion	R^2	RMSE (mg/kg)
90%	0.84	0.80
70%	0.82	0.85
50%	0.80	0.89
30%	0.78	0.94

3.4. Variable Importance Analysis

The Random Forest (RF) model's feature importance analysis provided valuable insights into the relative contributions of different predictors. Among environmental variables, soil pH emerged as the most influential factor with a normalized importance score of 1.00, followed by the Normalized Difference Vegetation Index (NDVI) at 0.87 and soil organic carbon content at 0.76. These findings align well with established geochemical principles, as soil pH directly affects arsenic mobility and bioavailability, while vegetation indices like the NDVI serve as proxies for plant health and potential contamination uptake.

Spatial features also played a significant role in the model's predictive performance. Ordinary kriging (OK)-derived spatial regionalization variables (SRs) exhibited the highest importance score (0.92) among spatial predictors, outperforming both spatial autocorrelation measures (SAs, 0.68) and SRs generated through other interpolation methods. This result highlights the effectiveness of geostatistical approaches in capturing spatial dependencies for heavy metal prediction.

Prior to finalizing the model, a comprehensive variable selection process was implemented to ensure robustness. The methodology began with an initial screening using variance inflation factors (VIF < 5) to eliminate multicollinear predictors. This was followed by recursive feature elimination with cross-validation and permutation importance testing within the RF framework. Although some variables, such as distance to roads (importance = 0.23), showed relatively minor contributions, they were retained due to their non-zero importance scores and potential ecological significance. Notably, the top three environmental predictors (soil pH, NDVI, and organic carbon) collectively accounted for 68% of the total feature importance, demonstrating their dominant role in explaining arsenic distribution patterns.

The strong performance of OK-based SRs further emphasizes the value of incorporating geostatistical interpolation techniques when engineering spatial features for environmental modeling. These findings not only validate the theoretical basis for variable selection but also provide practical guidance for future studies aiming to predict soil heavy metal contamination using similar methodologies.

Figure 3 compares model predictions against measured values, showing DF21's tighter clustering along the 1:1 line. The residual plots (insets) confirm its superior handling of high-concentration samples (>5 mg/kg), where PLSR and RF exhibited systematic underestimation.

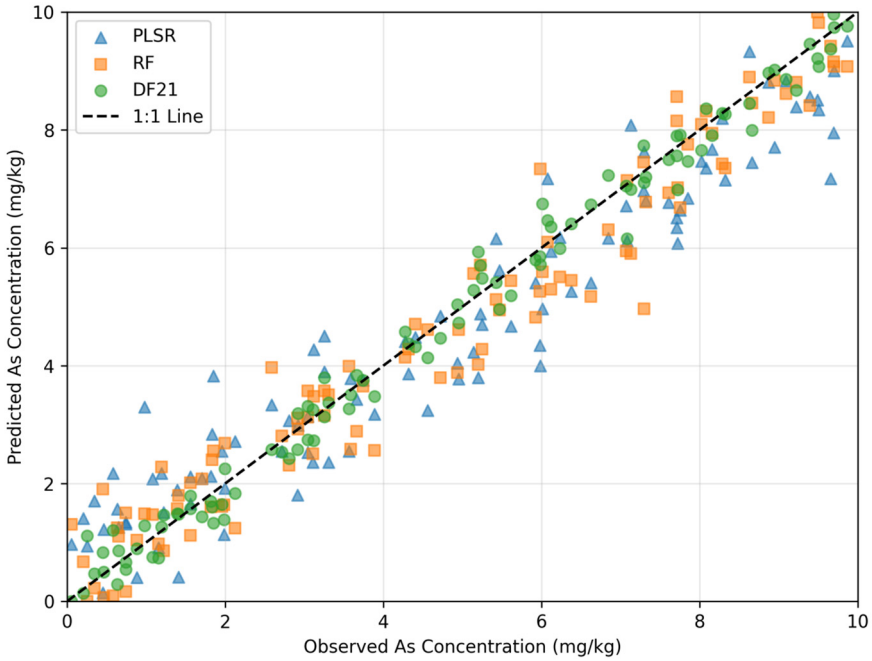

Figure 3. Observed vs. predicted arsenic concentrations for all models.

3.5. Cross-Region Validation

The framework's transferability was rigorously evaluated by applying the trained DF21 model with full spatial features (S-3 variables) to an independent dataset from Guangdong Province, China. This test region shares similar mining-induced contamination patterns with the original study area in Chenzhou but exhibits distinct soil characteristics and climatic conditions. Without any model retraining or parameter adjustments, the framework demonstrated robust cross-region performance, achieving an R^2 of 0.71–0.72 and RMSE of 0.91–0.93 mg/kg.

While these results represent a 13–14% decrease in predictive accuracy compared to the Chenzhou application (where R^2 reached 0.85), they significantly outperform conventional geostatistical methods. Traditional kriging approaches typically yield R^2 values between 0.58–0.62 when applied to new regions, highlighting the superior generalization capability of our integrated framework. The moderate performance decline observed can be attributed to differences in local soil properties and environmental conditions between the two regions.

This validation confirms that the framework maintains reasonable predictive accuracy when transferred to areas with broadly similar contamination sources (mining activities) and landscape characteristics, despite variations in specific soil parameters. The results suggest that the spatial feature engineering approach provides meaningful transfer learning benefits compared to conventional methods, particularly for regions sharing comparable pollution mechanisms but differing in secondary environmental factors. The successful cross-region application underscores the framework's potential for broader geographical implementation in mining-affected areas across China.

3.6. Extrapolation Testing Protocol

We evaluated extrapolation capability by (1) excluding all samples from three western townships during training, then testing exclusively on these unseen areas and (2) artificially generating extreme As concentrations ($\pm 3\sigma$) in 10% of the test data. The model maintained an $R^2 > 0.70$ in both scenarios, with slightly elevated but acceptable RMSE (0.95–1.02 mg/kg).

4. Discussion

This study advances soil heavy metal prediction by addressing three key challenges: the opaque treatment of spatial features in machine learning through systematic variable testing, scale-dependent spatial patterns via hierarchical SR construction, and data scarcity issues by maintaining accuracy with limited training samples. The 28–35% R^2 improvement using full spatial features quantitatively validates these methodological breakthroughs.

4.1. Limitations and Robustness of Variable Combinations and Models

Several important limitations should be considered when implementing this framework. First, while our spatial regionalization approach demonstrates robustness with limited training data, its performance remains dependent on the spatial distribution of sampling points. In areas with extreme sampling sparsity or clustered sampling designs, the interpolation-based SRs may introduce artifacts that propagate through the prediction pipeline [45]. Field validation in such scenarios would be particularly valuable. The framework currently treats environmental covariates and spatial features as independent inputs, potentially missing important interactions. For instance, soil pH—our most important environmental predictor—likely modifies both the mobility of arsenic and its spatial distribution patterns [46]. Future implementations could benefit from explicitly modeling these interactions through feature engineering or hierarchical modeling approaches.

The consistent outperformance of DF21 within our SA + SR + EC framework finds support in the broader spatial prediction literature. Sergeev et al. [47] demonstrated similar advantages when combining spatial autocorrelation measures with ensemble machine learning for various soil contaminants, noting that Deep Forest architectures particularly excelled at capturing hierarchical spatial patterns. This aligns with our findings where DF21's multi-grained scanning capability effectively processed both local (SA) and regional (SR) spatial features. Veronesi et al. [48] further corroborate these results, showing that advanced ensemble methods outperform traditional geostatistical approaches when environmental covariates exhibit complex, non-linear relationships with target variables. Our framework extends these insights by demonstrating that explicitly modeling spatial hierarchies through SR variables provides additional predictive power beyond conventional spatial autocorrelation measures, particularly for extreme concentration values (>90th percentile) where traditional methods often underperform.

While the DF21 model achieved superior accuracy, its computational demands (~40% longer training time than RF) may constrain practical applications, particularly for (1) real-time monitoring scenarios requiring frequent model updates, (2) large-area assessments with high-resolution spatial data, and (3) resource-constrained settings with limited computing infrastructure. This trade-off between accuracy and computational efficiency should guide model selection for specific applications.

All models exhibited reduced performance at extreme arsenic concentrations (>90th percentile), with systematic underestimation observed above 5 mg/kg (Figure 3). This suggests the need for specialized approaches like (1) quantile regression for extreme value prediction, (2) two-stage modeling separating background and hotspot prediction, and (3) tailored loss functions that penalize extreme value errors more heavily [49].

The reported model performance metrics (R^2 = 0.85, RMSE = 0.78 mg/kg), while seemingly modest, mark a substantial advancement over existing approaches. Compared to traditional geostatistical methods that achieved R^2 values of 0.58–0.65 in similar Chenzhou-area studies and environmental-variable-only baseline models (R^2 = 0.68 in our S-1 test), the current framework demonstrates meaningful progress. The 28–35% relative improvement from spatial feature integration proves particularly valuable for real-world applications where soil heterogeneity inherently limits absolute accuracy. This robustness is further

evidenced by subset analyses showing maintained performance ($R^2 > 0.78$) with as little as 30% training data.

The framework's transferability was systematically evaluated through external validation with Guangdong datasets. Results confirm its predictive power remains strong when applied to regions sharing key characteristics with the original study area: similar contamination sources (mining activities), comparable landscape features (hilly terrain with rice paddies), and analogous monitoring protocols. However, performance notably declines when applied to fundamentally different environments like urban soils or desert regions. This underscores how domain similarity governs the effectiveness of transfer learning approaches in geospatial contamination modeling. The consistent pattern of degradation across dissimilar environments suggests that while the framework advances generalizability, it remains bounded by the physical and anthropogenic contexts embedded in its training data.

4.2. Practical Applications in Environmental Monitoring and Policy

From an applied perspective, the framework offers tangible benefits for regulatory decision-making. The spatial explicitness of SRs enables targeted identification of contamination hotspots, optimizing remediation efforts and monitoring budgets. For example, Figure 2 reveals localized As accumulation near hydrological convergence zones—a pattern consistent with metal transport mechanisms that traditional sampling grids might miss [50].

The modular design also facilitates integration with existing environmental databases. By encoding spatial structures as model inputs rather than hard-coded assumptions, the approach can adapt to diverse regions without requiring algorithmic modifications. This flexibility is critical for scaling to national or global monitoring initiatives, where soil heterogeneity and data availability vary widely [51].

While the reported R^2 (0.85) and RMSE (0.78 mg/kg) may appear modest, they represent significant improvements over traditional geostatistical methods (R^2 = 0.58–0.65 in comparable Chenzhou-area studies) and environmental-variable-only models (R^2 = 0.68 in our S-1 baseline). More importantly, the 28–35% relative improvement from spatial feature integration (Table 2) demonstrates the framework's value for practical applications where absolute accuracy is constrained by inherent soil heterogeneity. This is particularly relevant given that our subset analyses show the method maintains an $R^2 > 0.78$ even with only 30% training data.

For regulatory applications, we recommend the following: 1. Priority implementation: OK-derived SRs given their consistent performance (normalized importance = 0.92). 2. Cost-effective monitoring: Use variable importance rankings (Figure 3) to guide field sampling, with the primary focus being pH (importance = 1.00) and NDVI (0.87) measurements and the secondary focus being organic carbon (0.76) in budget-constrained areas. 3. Phased adoption: the initial stage comprises EC + SA combinations (R^2 = 0.75) and the advanced stage full SR integration (R^2 = 0.85), as resources permit.

However, policy adoption faces barriers related to interpretability. While RF and DF21 provide feature importance metrics, their "black-box" nature complicates communication with stakeholders who prioritize causal understanding over predictive accuracy. Developing hybrid models that combine machine learning with mechanistic soil chemistry principles could bridge this gap [52].

4.3. Future Directions: Dynamic Data Integration and Model Interpretability

Advancing the methodology in this field can be guided by three key directions. Dynamic spatial modeling is one such direction. Presently, spatial representations (SRs) are static snapshots, failing to capture the evolving nature of heavy metal distributions, which

are influenced by climatic, hydrological, and anthropogenic factors. To create more accurate and up-to-date models, it is essential to incorporate time-series remote sensing data and mechanistic transport models. This approach would allow for dynamic updates of spatial features, as proposed in reference [53].

Another important direction is uncertainty quantification. The current framework provides only point estimates without confidence intervals, which limits its applicability in decision-making processes. To address this limitation, Bayesian approaches or ensemble methods could be utilized. These methods would propagate uncertainty from interpolation, variable selection, and model parameters, thereby supporting more reliable risk-based decision-making, as discussed in reference [54].

The third direction is cross-domain transfer learning. By leveraging pre-trained models from data-rich regions, deployment in understudied areas could be accelerated. However, this approach faces challenges related to domain shifts. Techniques such as adversarial adaptation or spatial covariance matching can be employed to address these challenges, as highlighted in reference [55].

Recent methodological advances offer promising directions for extending this framework. The graph neural network approach developed by Wang et al. [56] demonstrates how explicitly modeling spatial relationships as graph edges can improve heavy metal prediction, suggesting potential synergies with our regionalization approach. Similarly, Li et al.'s [57] work on interpretable machine learning for soil contaminants could help bridge the gap between our framework's predictive accuracy and the need for explainable results in policy applications. Future iterations might combine these approaches while addressing our current limitations in dynamic modeling and uncertainty quantification.

Future work should also explore the framework's applicability to other contaminants (e.g., cadmium, lead) and ecosystems (e.g., wetlands, urban soils), where spatial dependencies may follow distinct patterns. Comparative studies with emerging techniques like graph neural networks or physics-informed machine learning would further validate its relative advantages [58].

5. Conclusions

The study presents a robust framework for predicting soil heavy metal content by integrating environmental covariates with spatial autocorrelation and regionalization variables. The results demonstrate that spatial features significantly enhance prediction accuracy, with the DF21 model achieving superior performance (R^2 = 0.85) when all variable types are combined. The hierarchical evaluation of variable sets confirms that spatial regionalization variables (SRs) derived from interpolation methods, particularly ordinary kriging, contribute substantially to model improvement.

The framework's modularity allows for flexible adaptation across different geographical contexts and data availability scenarios. Subset analyses reveal its resilience to reduced training data, making it suitable for regions with sparse monitoring networks. The dominance of soil pH and NDVI among environmental predictors aligns with established geochemical principles, while the strong performance of SRs underscores the importance of explicitly modeling spatial dependencies.

Practical applications include targeted contamination monitoring and optimized remediation strategies, supported by the framework's ability to identify localized hotspots. Future research should focus on dynamic spatial modeling, uncertainty quantification, and cross-domain transferability to further enhance predictive capabilities. By bridging the gap between environmental and spatial data, this approach offers a scalable solution for soil contamination assessment and sustainable land management.

Author Contributions: Conceptualization, X.C., H.Z. and C.U.I.W.; Data curation, X.C. and Z.S.; Formal analysis, X.C., Z.S., H.Z. and C.U.I.W.; Methodology, X.C. and H.Z.; Software, X.C.; validation, X.C. and H.Z.; Investigation, X.C. and H.Z.; Writing—original draft, X.C. and H.Z.; Writing—review and editing, X.C., H.Z., Z.S. and C.U.I.W.; visualization, X.C. and C.U.I.W.; supervision, X.C., H.Z. and C.U.I.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

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

References

1. Xiang, M.; Li, Y.; Yang, J.; Lei, K.; Li, Y.; Li, F.; Zheng, D.; Fang, X.; Cao, Y. Heavy Metal Contamination Risk Assessment and Correlation Analysis of Heavy Metal Contents in Soil and Crops. *Environ. Pollut.* **2021**, *278*, 116911. [CrossRef] [PubMed]
2. Chen, X.; Zhang, H.; Wong, C.U.I.; Li, F.; Xie, S. Assessment of Heavy Metal Contamination and Ecological Risk in Soil within the Zheng–Bian–Luo Urban Agglomeration. *Processes* **2024**, *12*, 996. [CrossRef]
3. Suruliandi, A.; Mariammal, G.; Raja, S. Crop Prediction Based on Soil and Environmental Characteristics Using Feature Selection Techniques. *Math. Comput. Model. Dyn. Syst.* **2021**, *27*, 117–140. [CrossRef]
4. Xie, Z.M.; Huang, C.Y. Control of Arsenic Toxicity in Rice Plants Grown on an Arsenic-Polluted Paddy Soil. *Commun. Soil Sci. Plant Anal.* **1998**, *29*, 2471–2477. [CrossRef]
5. Jia, X.; Cao, Y.; O'Connor, D.; Zhu, J.; Tsang, D.C.; Zou, B.; Hou, D. Mapping Soil Pollution by Using Drone Image Recognition and Machine Learning at an Arsenic-Contaminated Agricultural Field. *Environ. Pollut.* **2021**, *270*, 116281. [CrossRef]
6. Li, X.; Wang, H.; Qin, S.; Lin, L.; Wang, X.; Cornelis, W. Evaluating Ensemble Learning in Developing Pedotransfer Functions to Predict Soil Hydraulic Properties. *J. Hydrol.* **2024**, *640*, 131658. [CrossRef]
7. Lombard, N.; Prestat, E.; van Elsas, J.D.; Simonet, P. Soil-Specific Limitations for Access and Analysis of Soil Microbial Communities by Metagenomics. *FEMS Microbiol. Ecol.* **2011**, *78*, 31–49. [CrossRef]
8. Palansooriya, K.N.; Li, J.; Dissanayake, P.D.; Suvarna, M.; Li, L.; Yuan, X.; Sarkar, B.; Tsang, D.C.; Rinklebe, J.; Wang, X.; et al. Prediction of Soil Heavy Metal Immobilization by Biochar Using Machine Learning. *Environ. Sci. Technol.* **2022**, *56*, 4187–4198. [CrossRef]
9. Jingzhe, W.; Jianing, Z.; Weifang, H.; Songchao, C.; Ivan, L.; Mojtaba, Z.; Xiaodong, Y. Remote Sensing of Soil Degradation: Progress and Perspective. *Int. Soil Water Conserv. Res.* **2023**, *11*, 429–454.
10. Liu, Z.; Lu, Y.; Peng, Y.; Zhao, L.; Wang, G.; Hu, Y. Estimation of Soil Heavy Metal Content Using Hyperspectral Data. *Remote Sens.* **2019**, *11*, 1464. [CrossRef]
11. Chen, X.; Cui, F.; Wong, C.; Zhang, H.; Wang, F. An Investigation into the Response of the Soil Ecological Environment to Tourist Disturbance in Baligou. *PeerJ* **2023**, *9*, E15780. [CrossRef] [PubMed]
12. Wang, H.; Zhao, M.; Huang, X.; Song, X.; Cai, B.; Tang, R.; Sun, J.; Han, Z.; Yang, J.; Liu, Y.; et al. Improving Prediction of Soil Heavy Metal (Loid) Concentration by Developing a Combined Co-Kriging and Geographically and Temporally Weighted Regression (GTWR) Model. *J. Hazard. Mater.* **2024**, *468*, 133745. [CrossRef] [PubMed]
13. Feng, C.; Yee, L.; ChangLin, M.; Tung, F. Backfitting Estimation for Geographically Weighted Regression Models with Spatial Autocorrelation in the Response. *Geogr. Anal.* **2021**, *54*, 357–381.
14. Zheng, Y.; Zhang, G.; Tan, S.; Feng, L. Research on Progress of Forest Fire Monitoring with Satellite Remote Sensing. *Agric. Rural. Stud.* **2023**, *1*, 0008. [CrossRef]
15. Anthony, T. Assessment of Heavy Metal Contamination in Wetlands Soils Around an Industrial Area Using Combined GIS-Based Pollution Indices and Remote Sensing Techniques. *Air Soil Water Res.* **2023**, *16*, 11786221231214062. [CrossRef]
16. Li, X. Influence of Variation of Soil Spatial Heterogeneity on Vegetation Restoration. *Sci. China Ser. D Earth Sci.* **2005**, *48*, 2020–2031. [CrossRef]
17. Song, I.; Kim, D. Three Common Machine Learning Algorithms Neither Enhance Prediction Accuracy Nor Reduce Spatial Autocorrelation in Residuals: An Analysis of Twenty-Five Socioeconomic Data Sets. *Geogr. Anal.* **2023**, *55*, 585–620. [CrossRef]
18. Li, Y.; Rahardjo, H.; Satyanaga, A.; Rangarajan, S.; Lee, D.T.T. Soil Database Development with the Application of Machine Learning Methods in Soil Properties Prediction. *Eng. Geol.* **2022**, *306*, 106769. [CrossRef]
19. Song, X.; Sun, Y.; Wang, H.; Huang, X.; Han, Z.; Shu, Y.; Wu, J.; Zhang, Z.; Zhong, Q.; Li, R.; et al. Uncovering Soil Heavy Metal Pollution Hotspots and Influencing Mechanisms through Machine Learning and Spatial Analysis. *Environ. Pollut.* **2025**, *370*, 125901. [CrossRef]

20. Yang, H.; Huang, K.; Zhang, K.; Weng, Q.; Zhang, H.; Wang, F. Predicting Heavy Metal Adsorption on Soil with Machine Learning and Mapping Global Distribution of Soil Adsorption Capacities. *Environ. Sci. Technol.* **2021**, *55*, 14316–14328. [CrossRef]
21. Hu, H.; Zhou, W.; Liu, X.; Guo, G.; He, Y.; Zhu, L.; Chen, D.; Miao, R. Machine Learning Combined with Geodetector to Predict the Spatial Distribution of Soil Heavy Metals in Mining Areas. *Sci. Total Environ.* **2025**, *959*, 178281. [CrossRef] [PubMed]
22. Wang, W.; Chen, S.; Chen, L.; Wang, L.; Chao, Y.; Shi, Z.; Lin, D.; Yang, K. Drivers Distinguishing of PAHs Heterogeneity in Surface Soil of China Using Deep Learning Coupled with Geo-Statistical Approach. *J. Hazard. Mater.* **2024**, *468*, 133840. [CrossRef] [PubMed]
23. Hua, W.; Junfeng, Z.; Fubao, Z.; Weiwei, Z. Analysis of Spatial Pattern of Aerosol Optical Depth and Affecting Factors Using Spatial Autocorrelation and Spatial Autoregressive Model. *Environ. Earth Sci.* **2016**, *75*, 822. [CrossRef]
24. Kuang, Y.; Chen, X. Spatial Heterogeneity of Forest Carbon Stocks in the Xiangjiang River Basin Urban Agglomeration: Analysis and Assessment Based on the Multiscale Geographically Weighted Regression (MGWR) Model. *Front. Environ. Sci.* **2025**, *13*, 1573438. [CrossRef]
25. Li, J.; Heap, A.D. Spatial Interpolation Methods Applied in the Environmental Sciences: A Review. *Environ. Model. Softw.* **2014**, *53*, 173–189. [CrossRef]
26. Lv, J. Multivariate Receptor Models and Robust Geostatistics to Estimate Source Apportionment of Heavy Metals in Soils. *Environ. Pollut.* **2019**, *244*, 72–83. [CrossRef]
27. Pauchard, A.; Alaback, P.B.; Edlund, E.G. Plant Invasions in Protected Areas at Multiple Scales: Linaria Vulgaris (Scrophulariaceae) in the West Yellowstone Area. *West. N. Am. Nat.* **2003**, *63*, 416–428.
28. Sreenivas, K.; Sujatha, G.; Sudhir, K.; Kiran, D.V.; Fyzee, M.; Ravisankar, T.; Dadhwal, V. Spatial Assessment of Soil Organic Carbon Density through Random Forests Based Imputation. *J. Indian Soc. Remote Sens.* **2014**, *42*, 577–587. [CrossRef]
29. Chen, X.; Zhang, H.; Wong, C.U.I. Spatial Distribution Characteristics and Pollution Evaluation of Soil Heavy Metals in Wulongdong National Forest Park. *Sci. Rep.* **2024**, *14*, 8880. [CrossRef]
30. Gadepalle, V.P.; Ouki, S.K.; Herwijnen, R.V.; Hutchings, T. Immobilization of Heavy Metals in Soil Using Natural and Waste Materials for Vegetation Establishment on Contaminated Sites. *Soil Sediment Contam.* **2007**, *16*, 233–251. [CrossRef]
31. Shu, X.; Gao, L.; Yang, J.; Xia, J.; Song, H.; Zhu, L.; Zhang, K.; Wu, L.; Pang, Z. Spatial Distribution Characteristics and Influencing Factors of Soil Organic Carbon Based on the Geographically Weighted Regression Model. *Environ. Monit. Assess.* **2024**, *196*, 1083. [CrossRef] [PubMed]
32. Tilahun, Y.; Xiao, Q.; Ashango, A.A.; Han, X.; Negewo, M. Prediction of Spatial Soil-California Bearing Ratio of Subgrade Soil Using Particle Swarm Optimization—Artificial Intelligence Method. *Transp. Infrastruct. Geotechnol.* **2025**, *12*, 80. [CrossRef]
33. Dai, X.; Wang, Z.; Liu, S.; Yao, Y.; Zhao, R.; Xiang, T.; Fu, T.; Feng, H.; Xiao, L.; Yang, X.; et al. Hyperspectral Imagery Reveals Large Spatial Variations of Heavy Metal Content in Agricultural Soil-A Case Study of Remote-Sensing Inversion Based on Orbita Hyperspectral Satellites (OHS) Imagery. *J. Clean. Prod.* **2022**, *380*, 134878. [CrossRef]
34. Galelli, S.; Humphrey, G.B.; Maier, H.R.; Castelletti, A.; Dandy, G.C.; Gibbs, M.S. An Evaluation Framework for Input Variable Selection Algorithms for Environmental Data-Driven Models. *Environ. Model. Softw.* **2014**, *62*, 33–51. [CrossRef]
35. Zhang, Y.; Lei, M.; Li, K.; Ju, T. Spatial Prediction of Soil Contamination Based on Machine Learning: A Review. *Front. Environ. Sci. Eng.* **2023**, *17*, 93. [CrossRef]
36. Wang, D.; Wang, M.; Qiao, X. Support Vector Machines Regression and Modeling of Greenhouse Environment. *Comput. Electron. Agric.* **2008**, *66*, 46–52. [CrossRef]
37. Zhao, M.; Wang, H.; Sun, J.; Tang, R.; Cai, B.; Song, X.; Huang, X.; Huang, J.; Fan, Z. Spatio-Temporal Characteristics of Soil Cd Pollution and Its Influencing Factors: A Geographically and Temporally Weighted Regression (GTWR) Method. *J. Hazard. Mater.* **2023**, *446*, 130613. [CrossRef]
38. Kuang, Y.; Chen, X.; Zhu, C. Characteristics of Soil Heavy Metal Pollution and Health Risks in Chenzhou City. *Processes* **2024**, *12*, 623. [CrossRef]
39. Chen, S.; Li, B.; Cao, J.; Mao, B. Research on Agricultural Environment Prediction Based on Deep Learning. *Procedia Comput. Sci.* **2018**, *139*, 33–40. [CrossRef]
40. Sun, Y.; Chen, S.; Jiang, H.; Qin, B.; Li, D.; Jia, K.; Wang, C. Towards Interpretable Machine Learning for Observational Quantification of Soil Heavy Metal Concentrations under Environmental Constraints. *Sci. Total Environ.* **2024**, *926*, 171931. [CrossRef]
41. Zhai, L.; Liao, X.; Chen, T.; Yan, X.; Xie, H.; Wu, B.; Wang, L. Regional Assessment of Cadmium Pollution in Agricultural Lands and the Potential Health Risk Related to Intensive Mining Activities: A Case Study in Chenzhou City, China. *J. Environ. Sci.* **2008**, *20*, 696–703. [CrossRef] [PubMed]
42. 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]
43. Wei, S.; Dai, Y.; Liu, B.; Zhu, A.; Duan, Q.; Wu, L.; Ji, D.; Ye, A.; Yuan, H.; Zhang, Q.; et al. A China Data Set of Soil Properties for Land Surface Modeling. *J. Adv. Model. Earth Syst.* **2013**, *5*, 212–224.

44. Sun, Q.; Miao, C.; Duan, Q.; Kong, D.; Ye, A.; Di, Z.; Gong, W. Would the 'Real' Observed Dataset Stand up? A Critical Examination of Eight Observed Gridded Climate Datasets for China. *Environ. Res. Lett.* **2014**, *9*, 015001. [CrossRef]
45. Miao, S.; Ni, G.; Kong, G.; Yuan, X.; Liu, C.; Shen, X.; Gao, W. A Spatial Interpolation Method Based on 3D-CNN for Soil Petroleum Hydrocarbon Pollution. *PLoS ONE* **2025**, *20*, e0316940. [CrossRef]
46. Justyna, K.; Janusz, P. Temporal and Spatial Variations of Selected Biomarker Activities in Flounder (Platichthys Flesus) Collected in the Baltic Proper. *Ecotoxicol. Environ. Saf.* **2008**, *70*, 379–391.
47. Sergeev, A.P.; Buevich, A.G.; Baglaeva, E.M.; Shichkin, A.V. Combining Spatial Autocorrelation with Machine Learning Increases Prediction Accuracy of Soil Heavy Metals. *Catena* **2019**, *174*, 425–435. [CrossRef]
48. Veronesi, F.; Schillaci, C. Comparison between Geostatistical and Machine Learning Models as Predictors of Topsoil Organic Carbon with a Focus on Local Uncertainty Estimation. *Ecol. Indic.* **2019**, *101*, 1032–1044. [CrossRef]
49. Tian, Y.; Su, D.; Lauria, S.; Liu, X. Recent Advances on Loss Functions in Deep Learning for Computer Vision. *Neurocomputing* **2022**, *497*, 129–158. [CrossRef]
50. Babu, G.R.; Gokuldhev, M.; Brahmanandam, P.S. Integrating IoT for Soil Monitoring and Hybrid Machine Learning in Predicting Tomato Crop Disease in a Typical South India Station. *Sensors* **2024**, *24*, 6177. [CrossRef]
51. Amato, F.; Guignard, F.; Robert, S.; Kanevski, M. A Novel Framework for Spatio-Temporal Prediction of Environmental Data Using Deep Learning. *Sci. Rep.* **2020**, *10*, 22243. [CrossRef] [PubMed]
52. Huang, J.; Fan, G.; Liu, C.; Zhou, D. Predicting Soil Available Cadmium by Machine Learning Based on Soil Properties. *J. Hazard. Mater.* **2023**, *460*, 132327. [CrossRef] [PubMed]
53. Eisenberg, J.N.; Bennett, D.H.; McKone, T.E. Chemical Dynamics of Persistent Organic Pollutants: A Sensitivity Analysis Relating Soil Concentration Levels to Atmospheric Emissions. *Environ. Sci. Technol.* **1998**, *32*, 115–123. [CrossRef]
54. Laura, U.; Moustapha, S.M.; MarcAndré, G.; Philipp, H.; Martin, S. Quantification of Conceptual Model Uncertainty in the Modeling of Wet Deposited Atmospheric Pollutants. *Risk Anal. Off. Publ. Soc. Risk Anal.* **2021**, *42*, 757–769.
55. Murakami, D.; Kajita, M.; Kajita, S. Spatial Process-Based Transfer Learning for Prediction Problems. *J. Geogr. Syst.* **2025**, *27*, 147–166. [CrossRef]
56. Wang, F.; Huo, L.; Li, Y.; Wu, L.; Zhang, Y.; Shi, G.; An, Y. A Hybrid Framework for Delineating the Migration Route of Soil Heavy Metal Pollution by Heavy Metal Similarity Calculation and Machine Learning Method. *Sci. Total Environ.* **2023**, *858*, 160065. [CrossRef]
57. Li, P.; Hao, H.; Mao, X.; Xu, J.; Lv, Y.; Chen, W.; Ge, D.; Zhang, Z. Convolutional Neural Network-Based Applied Research on the Enrichment of Heavy Metals in the Soil–Rice System in China. *Environ. Sci. Pollut. Res.* **2022**, *29*, 53642–53655. [CrossRef]
58. Zha, Y.; Yang, Y. Innovative Graph Neural Network Approach for Predicting Soil Heavy Metal Pollution in the Pearl River Basin, China. *Sci. Rep.* **2024**, *14*, 16505. [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.