agriculture

MDPI

*Article*

# Integrating Multi-Source Environmental Data to Identify Drivers and Assess Risks of Heavy Metal Contamination in Agricultural Soils

**Jie Wang [1,†], Chunli Zhao [2,†], Yanan Chen [3], Feng Zhang [4,*], Jiquan Zhang [5,*], Hansheng Liu [4] and Guzailinuer Aihemaitijiang [5]**

[1] College of Engineering and Technology, Jilin Agricultural University, Changchun 130118, China; 20252257@mails.jlau.edu.cn
[2] College of Forestry and Grassland Science, Jilin Agricultural University, Changchun 130118, China; zhaochunli@jlau.edu.cn
[3] College of Jilin Emergency Management, Changchun Institute of Technology, Changchun 130012, China; chenyn061@nenu.edu.cn
[4] College of Resources and Environment, Jilin Agricultural University, Changchun 130118, China; liuhansheng@mails.jlau.edu.cn
[5] School of Environment, Northeast Normal University, Changchun 130024, China; guzailinueraihemaitijiang999@nenu.edu.cn
* Correspondence: fengz@jlau.edu.cn (F.Z.); zhangjq022@nenu.edu.cn (J.Z.)
† These authors contributed equally to this wok.

**Abstract**

The pronounced spatial heterogeneity of soil heavy metals, coupled with the high costs and intensive resource requirements associated with large-scale sampling, presents substantial challenges for accurate spatial prediction and risk assessment. Although conventional machine learning methods demonstrate strong predictive capabilities, they often struggle to simultaneously address spatial non-stationarity, complex nonlinear relationships, and limited model interpretability. By integrating multi-source datasets with multiple machine learning models, SHAP analysis, Monte Carlo simulations, and bivariate Moran's I analysis, this study developed an integrated and interpretable framework for spatial prediction and risk assessment. The results revealed that (1) the proposed framework achieved high predictive accuracy, with coefficients of determination ($R^2$) exceeding 0.80 for arsenic (As), cadmium (Cd), copper (Cu), lead (Pb), chromium (Cr), and zinc (Zn). Climatic conditions and soil properties were identified as the dominant factors controlling the spatial distributions of heavy metals; (2) As and Cr were identified as the major contributors to carcinogenic risks, whereas children exhibited higher overall health risks than adults throughout the study area; (3) spatial hotspot analysis revealed significant heavy metal pollution hotspots in Gongzhuling City, northern Yitong County, and Dongliao County. This study provides new insights into soil heavy metal exposure and risk assessment, offering scientific support for targeted environmental management and pollution prevention strategies.

**Keywords:** soil heavy metals; explainable machine learning; SHAP interaction; multi-source data; risk assessment

Academic Editor: Yijun Yao

Received: 29 July 2026
Revised: 31 October 2026
Accepted: 1 November 2026
Published: 4 November 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) license.

## 1. Introduction

Soil underpins the production of approximately 95% of all food consumed by humanity [1]. Globally, an estimated 14–17% of agricultural topsoil exceeds the alert threshold for at least one toxic metal [2]. Over the past decade, between 38.2% and 84.1% of agricultural soils distributed across 93 cities within 30 provinces of China have exceeded permissible limits for eight heavy metals, including arsenic, mercury, and copper. With the rapid industrialisation and urbanisation in China, heavy metal enrichment in domestic soils has intensified markedly [3,4]. Owing to their toxicity and bioaccumulation, these elements significantly affect agriculture, ecological stability, and human health [5]. Although China has implemented comprehensive measures to mitigate heavy metal pollution—such as improving emission control systems, optimising industrial structures, strengthening regulatory oversight, and advancing pollution remediation [6–8] —heavy metal contamination remains a significant environmental challenge [9].

Current approaches for spatial prediction of soil heavy metals primarily include spatial modelling techniques and conventional machine learning algorithms, such as random forests (RFs) and support vector machines (SVMs) [10–13]. Although widely used in soil contamination research to enhance spatial predictive accuracy, many approaches fail to consider the spatial heterogeneity of heavy metals [14]. Geographically weighted regression (GWR) can capture spatial non-stationarity and reduce model errors caused by spatial heterogeneity. Nevertheless, their inherent black-box nature results in limited interpretability, reduced transparency, and challenges in generalisation [15]. The emergence of explainable artificial intelligence (XAI) techniques, exemplified by the Shapley Additive exPlanations (SHAP), has enabled quantitative analysis of feature contributions within machine learning models. These methods elucidate how input variables shape model behaviour, thereby improving transparency and supporting practical implementation and management [16].

Regression models such as GWR explicitly incorporate spatial effects, thereby improving the explanation of spatial variability and uncovering underlying distributional patterns [17]. Meanwhile, the eXtreme Gradient Boosting (XGBoost) model adopts a gradient boosting strategy that iteratively corrects residuals through multiple rounds of learning. Compared with the parallel ensemble structure of traditional models such as RF, it is more capable of capturing complex nonlinear relationships and achieves higher predictive accuracy for soil heavy metal data characterized by strong nonlinearity and interactions. The proposed framework enables robust predictive models to capture spatial heterogeneity and nonlinear complexities, yielding more precise estimates.

This study developed an integrated framework for identifying and managing soil heavy metal hotspots (Figure S1). Specifically, the proposed framework consists of the following components: (1) Integrate multi-source datasets, including field-sampled data, environmental covariates, and spatial covariates. (2) Develop a hybrid GWR-XGBoost model coupled with SHAP analysis to enhance model transparency and interpretability. (3) Analyze the spatial distribution patterns and driving factors of soil heavy metal contamination. (4) Conduct risk assessment and hotspot identification based on the predicted spatial distribution of soil heavy metals. This study establishes a conceptual framework for the spatial prediction and assessment of heavy metal contamination, providing new insights into environmental pollution prevention, control, and management.

## 2. Materials and Methods

### *2.1. Study Area*

The Liaohe River Basin of Jilin Province is located in southwestern Jilin Province and encompasses Gongzhuling City, as well as parts of Siping and Liaoyuan. The basin is an

https://doi.org/10.3390/agriculture16212416

important grain-producing region at both the national and provincial levels and a major centre for animal husbandry in Jilin Province. Cropland is the dominant land-use type in the basin, and agriculture plays a substantial role in the regional economy. Maize and other grain crops are widely cultivated, while livestock farming constitutes another major agricultural activity in the region. The basin also hosts industrial activities, including agricultural product processing, chemical manufacturing and automotive component production. Although these industries contribute to regional economic development, they add complexity to environmental management within the basin.

With the intensification of agricultural production and economic activities, some areas face environmental challenges, including agricultural non-point source pollution, inadequate collection and treatment of livestock manure, and industrial wastewater discharge. Fertilizer and pesticide application, agricultural drainage, and surface runoff can facilitate the transport of pollutants into surrounding environments and contribute to their accumulation. Industrial activities, domestic wastewater generation, and livestock farming may also contribute to localized environmental pressures. In this context, industrial emissions and atmospheric deposition, contaminated irrigation water, and certain agricultural inputs represent potential sources of heavy metals entering agricultural soils.

### *2.2. Data Collection*

#### 2.2.1. Sample Collection

The Liaohe River Basin of Jilin Province is located in the upper reaches of the Liaohe River Basin. The surficial geology is characterized by alluvial deposits and diluvial loess-like clay, underlain by sand and gravel, with Cretaceous strata at depth. The basin is predominantly covered by unconsolidated sediments, with bedrock exposed only in localized areas. The main soil types in the region include black soils, chernozems and meadow soils. The basin is characterized by a semi-humid, temperate monsoon climate, with four distinct seasons. Spring is generally dry and windy, whereas summer is warm and humid, with most precipitation occurring during this season. Mean annual precipitation is 545 mm, compared with a mean annual evaporation of 1020 mm.

A systematic sampling strategy was designed to collect agricultural soil samples across the Liaohe River Basin in Jilin Province, China (Figure 1). Representative maize field sites were selected along the basin from upstream to downstream to characterize regional patterns of soil heavy metal pollution. Eighty-two soil specimens were obtained using a five-point sampling design, with three replicates per location. Topsoil (0–20 cm) was collected with stainless steel shovels and sealed in polyethylene bags. Cr, Cu, Zn, and Pb concentrations were determined by flame atomic absorption spectrophotometry (HJ 491-2019). Concentrations were determined after microwave digestion (HJ 680-2013), whereas Cd concentrations were analysed using graphite furnace atomic absorption spectrophotometry (GB/T 17141-1997).

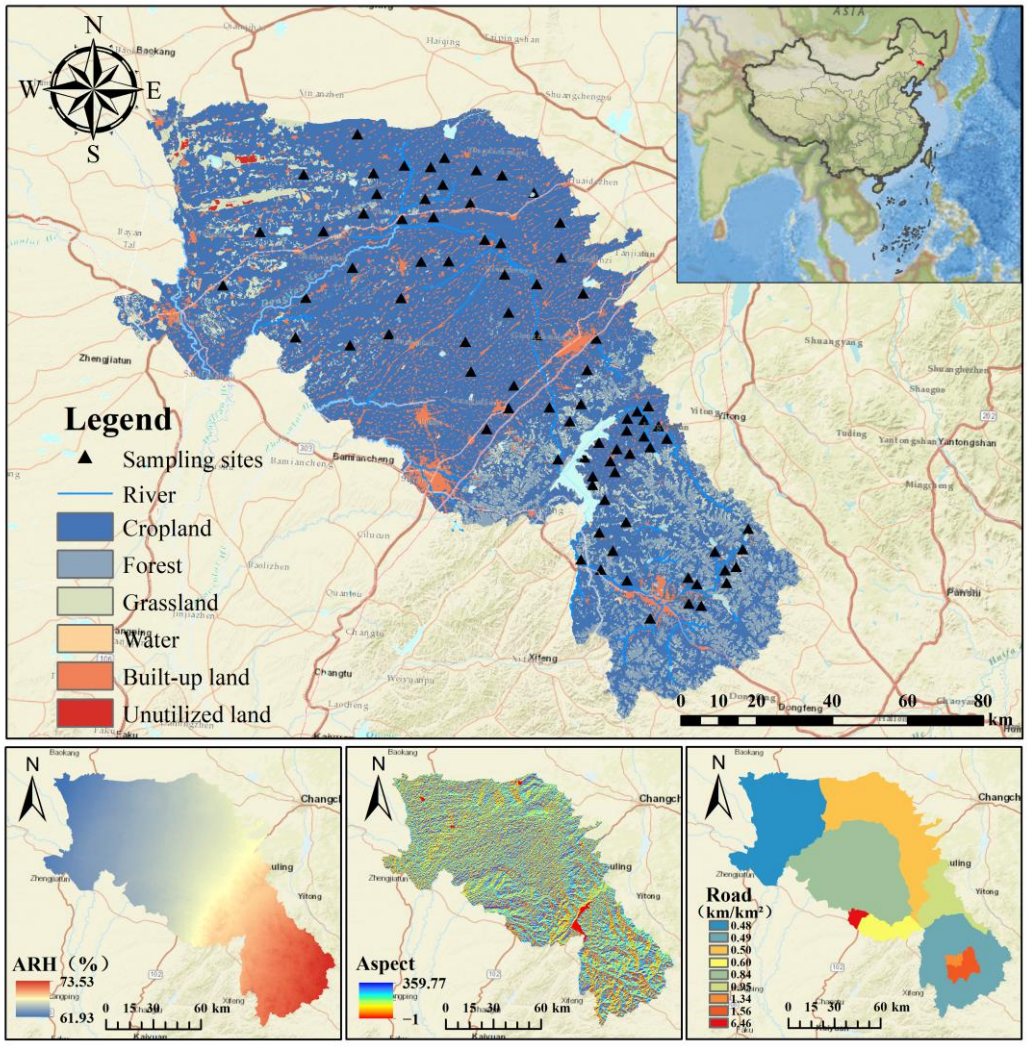



**Figure 1.** Schematic diagram of sampling points and three representative covariates in the study area. ARH—Annual relative humidity (%), Aspect—Slope aspect (°), Road—Density of road network (km/km²). (Background map data sources: Esri, HEPE, Garmin, USGS, Intermap, INCREMENTP, NRCan, Esri Japan, ME TI, Esri China (Hong Kong), Esri Korea, Esri (Thailand), NGCC, (c) OpenStreetMap contributors, and the GIS User Community).

### 2.2.2. Acquisition of Environmental Covariates

Multiple environmental drivers—including soil physicochemical attributes, climatic and ecological conditions, and human activities—jointly regulate heavy metal contamination. In this study, thirty-nine factors were selected and classified into six categories as environmental covariates (Table S1) (Figure 2). Sixteen soil physicochemical covariates—including exchangeable sodium (ES), exchangeable calcium (EC), available potassium (AK), available phosphorus (A-P), alkali-hydrolysable nitrogen (HN), soil pH (pH), cation exchange capacity (CEC), bulk density (BD), clay content (Clay), sand content (Sand), silt content (Silt), soil porosity (Por), soil organic matter (SOM), total potassium (TK), total nitrogen (TN), and total phosphorus (TP)—were selected as soil property covariates (Figure S2). These data were obtained from the National Tibetan Plateau Science Data Center (TPDC) [18] (https://data.tpdc.ac.cn/). Elevation (DEM), slope gradient (Slope), and aspect (Aspect) were employed as topographic conditions (Figure S3), with data obtained from the Geospatial Data Cloud (https://www.gscloud.cn/). Annual average wind speed (AWS) and five climatic covariates—annual average precipitation (AP), annual average temperature (AT), annual average relative humidity (ARH), and annual evaporation (EVA) —

were derived from the Resource and Environmental Science Data Platform (https://www.resdc.cn/) (Figure S4). Ecological functions included soil type (Soil Type), root abundance (RA), foliar area cover (FVC), normalised difference vegetation index (NDVI), and soil erosion (Erosion) (Figure S5). The social economy comprised population density (POP), road density (Road), railway density (Railway) and gross domestic product (GDP) (Figure S6), with datasets obtained from OpenStreetMap (https://www.openstreetmap.org/) and Scientific Data [19] (https://www.nature.com/sdata/ (accessed on 24 July 2025)). Industrial activity covariates included the chemical plant core density (CP), garment and textile factory core density (GTF), metal manufacturing plant core density (MMP), printing and dyeing factory core density (PDF), electronics factory core density (ETF), and all factory core density (ALL) (Figure S7). Corresponding industrial data were obtained from Guihuayun (https://guihuayun.com/).

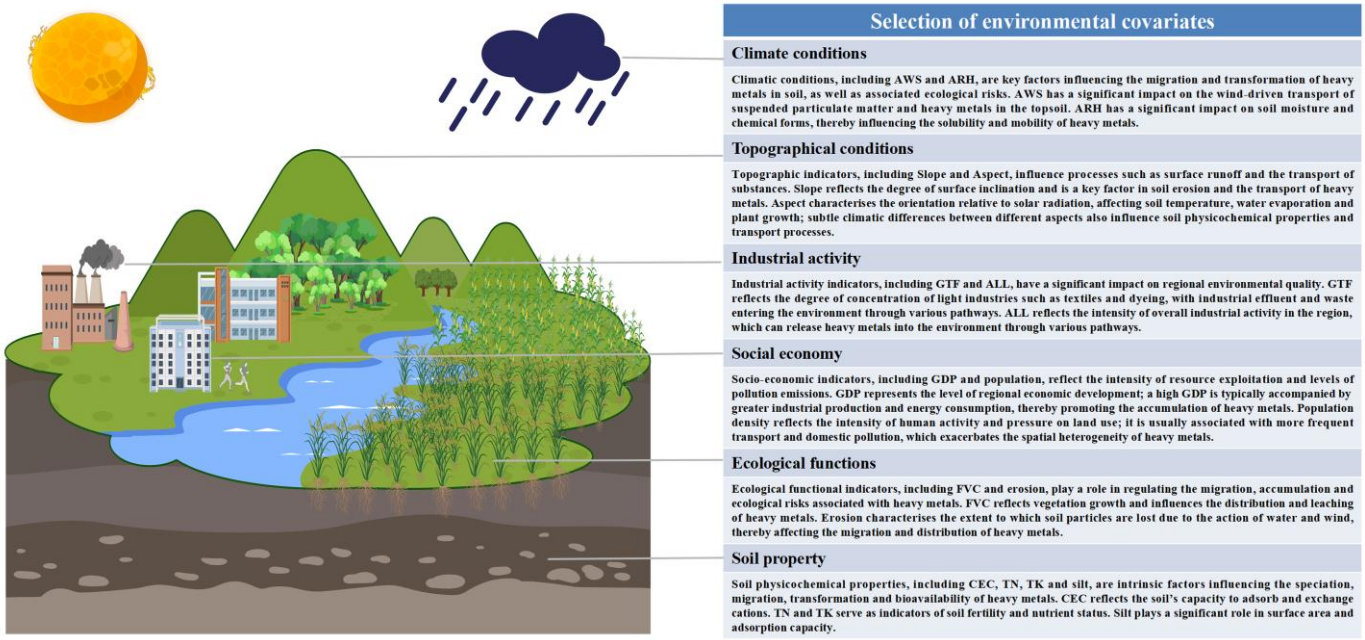



**Figure 2.** A description of the characteristics of environmental covariates. The diagram was created using BioGDP.com.

#### 2.2.3. Collection of Spatial Covariates

To assess how spatial distribution patterns affect soil heavy metal contamination, this study employed spatial regionalization covariates (SRs) (Text S1) [20]. This method transforms sparse, point-based data into continuous spatial layers, enhancing the representation of distribution patterns [21,22]. Moreover, incorporating spatial covariates into neural network models designed for high-dimensional, complex datasets markedly improves prediction robustness and reliability.

### *2.3. Data Preprocessing*

The original data were standardized using StandardScaler to ensure consistency across feature dimensions. Spearman's rank correlation analysis was applied for feature selection by calculating pairwise Spearman correlation coefficients with a threshold of 0.75. Two spatial features derived from GWR (spatial trend predictions and local residuals) were retained to prevent the loss of spatial information.

### *2.4. Modeling Methods*

#### 2.4.1. GWR-XGBoost Ensemble Model

This study develops a GWR-XGBoost hybrid model capable of capturing spatial heterogeneity in heavy metals and complex nonlinear relationships among features, improving model reliability and predictive accuracy. Spatial distribution patterns of soil heavy metals typically exhibit pronounced regional heterogeneity [23]. GWR represents a localized regression approach aimed at capturing spatial heterogeneity in variable associations [24]. On this basis, XGBoost was introduced to capture complex nonlinear relationships not adequately represented by GWR, with the spatial trend and residual components derived from GWR predictions incorporated as input features to further explore the response patterns of multi-source data. This integrated framework provides a more targeted modeling approach for predicting pollutants in complex spatial scenarios.

#### 2.4.2. Data Partitioning and Parameter Optimization

In this study, a spatially blocked five-fold cross-validation strategy was employed for data partitioning and model performance evaluation. After initial partitioning via K-means clustering, iterative refinement was performed using the adjacency matrix. The data were thus divided into five spatially independent blocks, enabling leakage-free five-fold cross-validation. This ensured no spatially adjacent samples were shared between training and test sets, providing an unbiased assessment of the model's spatial extrapolation capability.

Within this framework, hyperparameter optimization was conducted by minimizing RMSE while jointly considering $R^2$ as the performance evaluation metric. For the GWR module, the median spherical distance derived from randomly sampled points was used as the initial bandwidth. The bandwidth was varied within 0.5–1.5 times this value (five candidate values), and the optimal bandwidth was determined according to the maximum cross-validated $R^2$ (CV-$R^2$). For the XGBoost module, an overfitting-prevention parameter grid was defined considering the small sample size. The parameter min_child_weight was fixed at 8 to prevent ineffective node splits. All parameter combinations were evaluated in conjunction with spatially blocked cross-validation results. The optimal parameters for each fold were first determined, and the mode of these fold-specific optimal parameters was taken as the global optimal hyperparameter set.

#### 2.4.3. Model Evaluation

The RF is an ensemble learning method that mitigates overfitting and improves model generalization performance [25]. It constructs multiple decision trees and aggregates their predictions to improve prediction accuracy and robustness [26].

$$\mathrm{H}(\mathrm{x},\{\Theta_\mathrm{m}\}) = 1/\mathrm{M}\sum_{\mathrm{m}=1}^{\mathrm{M}} \mathrm{h_m}(\mathrm{x},\Theta_\mathrm{m}) \quad (1)$$

where $H(x,\{\Theta_m\})$ represents the overall prediction of the random forest, $h_m(x,\Theta_m)$ denotes the prediction of the $m$-th tree, $\{\Theta_m\}$ is the set of parameters for each tree, $M$ is the total number of trees.

XGBoost is a high-performance gradient boosting algorithm that forms a robust predictor by combining multiple weak learners [27]. During the iterative model training, each new decision tree is constructed to optimize the loss of all preceding trees [28]. This iterative process gradually improves the model and enhances prediction accuracy.

$$\mathrm{Obj} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{l}(\mathrm{yi},\hat{\mathrm{y}}_\mathrm{i}) + \sum_{\mathrm{i}=1}^{\mathrm{t}} \Omega(\mathrm{f_i}) \quad (2)$$

Here, $\sum_{i=1}^{n} l(yi, \hat{y}_i)$ represents the loss function term, where $n$ is the number of samples, $yi$ and $\hat{y}_i$ denote the observed and predicted values, respectively. $\sum_{i=1}^{t} \Omega(f_i)$ is the regularization term, which sums the complexity of all $t$ trees.

GWRF method extends the RF algorithm by integrating the GWR framework, forming a localized nonlinear machine learning model that captures regional variability and local dependencies in spatial data [29]. The model constructs an independent RF model for samples at each spatial location and incorporates geographic weighting to achieve more accurate local predictions [26].

$$\mathrm{Y_i} = \mathrm{a(u_i, v_i)x_i} + \mathrm{e} \tag{3}$$

Here, $a(u_i, v_i)x_i$ denotes the locally calibrated RF model prediction at location $(u_i, v_i)$, and $e$ represents the error term.

GWR is a spatial statistical technique developed to model relationships that vary across space. Brunsdon et al. (1996) first introduced the spatially varying coefficient model and proposed the GWR method, which has since become a fundamental tool for analyzing spatial heterogeneity [30]. Mathematical formulations for the weighted regression model and Gaussian kernel function are presented below:

$$\mathrm{y_i} = \beta_0(\mathrm{u_i}, \mathrm{v_i}) + \sum_{\mathrm{k=1}}^{\mathrm{p}} \beta_{\mathrm{k}}(\mathrm{u_i}, \mathrm{v_i})\mathrm{x_{ik}} + \varepsilon_{\mathrm{i}} \tag{4}$$

$y_i$ denotes the heavy metal concentration value at location $i$, $\beta_0(u_i, v_i)$ denotes the intercept term, $\beta_k(u_i, v_i)$ denotes the regression coefficient at position $(u_i, v_i)$, $x_{ik}$ denotes the value of the kth variable indicator at the $i$ th sampling point, $\varepsilon_i$ denotes the error term.

$$\mathrm{w_{ij}} = \exp\left(-1/2\left(\frac{\mathrm{d_{ij}}}{\mathrm{b}}\right)^2\right) \tag{5}$$

$d_{ij}$ denotes the distance between point i and point j, $b$ represents the bandwidth.

Evaluation metrics, including the coefficient of determination ($R^2$), root mean square error (RMSE), and mean absolute error (MAE), are used to reflect the predictive accuracy of the model.

$$\mathrm{R^2} = 1 - \frac{\sum_{\mathrm{i=1}}^{\mathrm{n}}(\hat{\mathrm{y}}_\mathrm{i} - \mathrm{y_i})^2}{\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{y_i} - \bar{\mathrm{y}})^2} \tag{6}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{\mathrm{i=1}}^{\mathrm{n}}(\hat{\mathrm{y}}_\mathrm{i} - \mathrm{y_i})^2}{\mathrm{n}}} \tag{7}$$

$$MAE = 1/n\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{8}$$

Here, $n$ denotes the number of samples, $y_i$ represents the observed concentration of the $i$-th sample, $\hat{y}_i$ is the model-predicted concentration, and $\bar{y}$ is the mean of all observed samples.

#### 2.4.4. Shapley Additive exPlanations (SHAP)

Shapley's method fairly distributes the incremental effects of model inputs on predictions, enabling assessment of feature importance and delivering consistent, transparent explanations [31].

$$\emptyset_i = \sum_{S \subseteq F \backslash \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!}\left[f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - \mathrm{fs}(x_S)\right] \tag{9}$$

$\emptyset_i$ denotes the SHAP value for feature I, $F$ denotes the set of all features, $S$ denotes the feature subset, and $fs(x_S)$ denotes the model's predicted value under this feature combination.

### *2.5. Risk Assessment*

The ecological risks associated with soil heavy metal contamination were assessed using the Potential Ecological Risk Index (PERI) method. The human health risk assessment framework recommended by the United States Environmental Protection Agency (USEPA) was employed to evaluate carcinogenic and non-carcinogenic risks to local adults and children through three exposure pathways: ingestion, inhalation, and dermal contact. Monte Carlo simulations were performed using Oracle Crystal Ball to characterize the cumulative probability distributions of health risks.

PERI method integrates soil heavy metal concentrations with toxicological characteristics by incorporating toxicity response coefficients into the assessment process.

$$C_f^i = \frac{C_i}{C_n^i} \tag{10}$$

$$E_r^i = T_r^i \times C_f^i \tag{11}$$

$$RI = \sum_{i=1}^{n} E_r^i \tag{12}$$

where $C_f^i$ is the contamination factor of heavy metal $i$; $C_i$ is the measured concentration (mg kg$^{-1}$); $C_n^i$ is the corresponding background concentration (mg kg$^{-1}$), obtained from the regional soil background survey of Jilin Province [32]; $T_r^i$ denotes the toxicity response coefficient of heavy metal $i$ (Table 1), adopted from Hakanson [33]; $E_r^i$ represents the potential ecological risk factor of heavy metal $i$; and $RI$ is the integrated potential ecological risk index for the study area. The detailed classification criteria are provided in Table 2.

**Table 1.** Toxicity response coefficients of heavy metals.

| Element | Pb | Cu | Zn | Cd | As | Cr |
|---|---|---|---|---|---|---|
| $T_r^i$ | 5 | 5 | 1 | 30 | 10 | 2 |

**Table 2.** Classification criteria for potential ecological risk.

| Level | Potential Ecological Risk Factor $E_r^i$ | Comprehensive Potential Ecological Risk Index *RI* | Potential Ecological Risk Level |
|---|---|---|---|
| 1 | $E_r^i < 40$ | $RI < 150$ | Low potential ecological risk |
| 2 | $40 \le E_r^i < 80$ | $150 \le RI < 300$ | Moderate potential ecological risk |
| 3 | $80 \le E_r^i < 160$ | $300 \le RI < 600$ | Considerable potential ecological risk |
| 4 | $160 \le E_r^i < 320$ | $600 \le RI < 1200$ | High potential ecological risk |
| 5 | $E_r^i \ge 320$ | $RI \ge 1200$ | Very high potential ecological risk |

Human health risk assessment was conducted using the heavy metal concentrations predicted by the model following the framework recommended by the USEPA.

$$ADD_{ing} = \frac{C_i \times IR_{ing} \times EF \times ED}{BW \times AT} \times 10^{-6} \tag{13}$$

$$ADD_{inh} = \frac{C_i \times IR_{inh} \times EF \times ED}{PEF \times BW \times AT} \tag{14}$$

$$ADD_{der} = \frac{C_i \times SA \times SL \times ABS \times EF \times ED}{BW \times AT} \times 10^{-6} \quad (15)$$

$$HQ_i = \frac{ADD_{ing,i}}{RfD_{ing,i}} + \frac{ADD_{inh,i}}{RfD_{inh,i}} + \frac{ADD_{der,i}}{RfD_{der,i}} \quad (16)$$

$$CR_i = ADD_{ing,i} \times SF_{ing,i} + ADD_{inh,i} \times SF_{inh,i} + ADD_{der,i} \times SF_{der,i} \quad (17)$$

$$THI = \sum_{i=1}^{n} HQ_i = \sum_{i=1}^{n} \frac{ADD_{ing,i}}{RfD_{ing,i}} + \sum_{i=1}^{n} \frac{ADD_{inh,i}}{RfD_{inh,i}} + \sum_{i=1}^{n} \frac{ADD_{der,i}}{RfD_{der,i}} \quad (18)$$

$$TCR = \sum_{i=1}^{n} CR_i = \sum_{i=1}^{n} ADD_{ing,i} \times SF_{ing,i} + \sum_{i=1}^{n} ADD_{inh,i} \times SF_{inh,i} + \sum_{i=1}^{n} ADD_{der,i} \times SF_{der,i} \quad (19)$$

where $ADD_{ing}$, $ADD_{inh}$, and $ADD_{der}$ represent the average daily doses of toxic elements through ingestion, inhalation, and dermal contact, respectively. The exposure parameters include ingestion rate (IngR), inhalation rate (InhR), exposure frequency (EF), exposure duration (ED), body weight (BW), and averaging time (AT). SA, SL, and ABS represent the exposed skin area, soil adherence factor, and dermal absorption factor, respectively, while PEF denotes the particle emission factor. SF denotes the slope factor, RfD is the reference dose (Table 3), $HQ_i$ represents the non-carcinogenic hazard quotient for the ith heavy metal, and $CR_i$ represents its carcinogenic risk. THI and TCR denote the total non-carcinogenic hazard index and total carcinogenic risk, respectively, calculated by summing the risk contributions from all heavy metals and exposure pathways. A TCR value greater than $1 \times 10^{-4}$ indicates an unacceptable carcinogenic risk. A TCR value between $1 \times 10^{-6}$ and $1 \times 10^{-4}$ is considered an acceptable carcinogenic risk level. When TCR is lower than $1 \times 10^{-6}$, the carcinogenic risk is considered negligible. A THI greater than 1 indicates potential non-carcinogenic effects, whereas a THI below 1 suggests that non-carcinogenic risks are negligible. In the absence of chromium speciation data, Cr(VI) specific toxicological parameters were conservatively applied to the measured total Cr concentrations for health risk assessment. The RfD and SF values for Cr(VI) were obtained from the literature [34]. The parameter settings used for the Monte Carlo simulation and cumulative probability distribution analysis are provided in Table 4.

**Table 3.** Reference doses (RfD) and slope factors (SF) of heavy metals.

| Heavy Metals | RfD | | | SF | | | Reference |
|---|---|---|---|---|---|---|---|
| | Ingest | Inhale | Dermal | Ingest | Inhale | Dermal | |
| Cr (VI) | 0.003 | 0.00003 | | 0.5 | 41 | | [34] |
| Cu | 0.04 | 0.04 | 0.012 | 1.7 | | 42.5 | [35] |
| Zn | 0.3 | 0.3 | 0.06 | | | | [35] |
| As | 0.0003 | 0.3 | 0.0001 | 0.5 | 42 | 20 | [35] |
| Cd | 0.001 | 0.001 | 0.00001 | 6.1 | 6.3 | 6.1 | [35] |
| Pb | 0.0035 | 0.0035 | 0.00053 | | | | [35] |

**Table 4.** Parameter settings for Monte Carlo simulation.

| Indices | Probabilistic Distribution | Unit | Adult | Children | Reference |
|---|---|---|---|---|---|
| IRing | Triangular | mg/d | TRI (4, 30, 52) | TRI (66, 103, 161) | [36] |
| IRinh | Log-normal | $m^3$/day | LN (16.57, 4.05) | LN (7.19, 1.62) | [37] |
| EF | Triangular | d/a | TRI (180,345.365) | TRI (180,345.365) | [35] |
| ED | Point | a | 24 | 6 | [36] |

| | | | | | |
|---|---|---|---|---|---|
| SA | Point | $cm^2$ | 5700 | 2800 | [35] |
| SL | Point | $mg/cm^2/d$ | 0.07 | 0.2 | [36] |
| ABS | Point | | 0.03 for As and 0.001 for other PHEs | 0.03 for As and 0.001 for other PHEs | [36] |
| PEF | Point | $m^3/kg$ | $1.36 \times 10^9$ | $1.36 \times 10^9$ | [36] |
| BW | Log-normal | kg | LN (57.59, 8.03) | LN (19.60, 1.96) | [38] |
| AT (Non-carcinogens) | Point | d | ED × 365 | ED × 365 | [36] |
| AT (carcinogens) | Point | d | 25,550 | 25,550 | [36] |

## 3. Results and Discussion

### *3.1. Soil Heavy Metal Contamination*

Field sampling in the Liaohe River Basin of Jilin Province, China, revealed mean concentrations of Cr, Cu, Zn, As, Cd, and Pb of 113.26 mg/kg, 18.83 mg/kg, 64.59 mg/kg, 9.17 mg/kg, 0.22 mg/kg, and 19.13 mg/kg, respectively (Figure S8). The mean concentrations of As, Cd, Cr, and Cu exceeded the corresponding provincial background values for Jilin soils, whereas those of Pb and Zn were below their respective background values (Table 5). The background-value exceedance rates, calculated as the percentages of soil samples with concentrations above the corresponding provincial background values, were 100% for both Cr and Cd, followed by As (73.17%), Cu (64.63%), Zn (31.71%), and Pb (20.73%). Pronounced spatial heterogeneity of heavy metal concentrations was observed, with Zn ranging from 34.50 to 80.83 mg/kg, corresponding to localised pollution hotspots [21]. The degree of variation further highlights disparities in their spatial distribution. For example, Cu and Cd exhibited variation coefficients of 0.25 and 0.24, likely reflecting the joint effects of natural processes and human activities [39].

**Table 5.** Concentrations of Heavy Metals in Cultivated Soils of the Liaohe River Basin, Jilin Province ($mg\ kg^{-1}$).

| Elements | Minimum | Maximum | Mean | Background Values [a] | Exceedance Rate (%) [b] | Risk Benchmark [c] |
|---|---|---|---|---|---|---|
| As | 7.05 | 11.18 | 9.17 | 8.38 | 73.17 | 25 |
| Cd | 0.13 | 0.30 | 0.22 | 0.11 | 100 | 0.6 |
| Cr | 71.93 | 165.25 | 113.26 | 45.71 | 100 | 250 |
| Cu | 10.41 | 27.58 | 18.83 | 16.24 | 64.63 | 100 |
| Pb | 11.10 | 27.33 | 19.13 | 23.08 | 20.73 | 170 |
| Zn | 34.50 | 80.83 | 64.59 | 71.68 | 31.71 | 300 |

[a] Background values: Study on the Soil Background Values of Jilin Province Conducted in Collaboration with the Jilin Environmental Monitoring Center, Jilin Environmental Protection Research Institute, and the Changchun Geographic Research Institute of the Chinese Academy of Sciences [32]. [b] Exceedance rate represents the percentage of soil samples with concentrations exceeding the corresponding provincial soil background values. [c] Risk benchmark: Risk Screening Values in Standards for Managing Soil Contamination Risks in Agricultural Soils Approved by the National Ecological and Environmental Protection Agency (GB 15618-2018).

### *3.2. Prediction of Soil Heavy Metal Concentrations*

Model predictive performance was systematically evaluated through $R^2$, RMSE, and MAE. As illustrated in Figure 3a, the GWR-XGBoost model exhibited a markedly superior fit compared with the standalone GWR model, GWRF models, and conventional machine learning algorithms such as XGBoost and RF. Taking Cu as an example, the $R^2$ value of the GWR-XGBoost model increased by 0.436 compared to GWR. Relative to the conventional models XGBoost and RF, it also improved by 0.123 and 0.324, respectively. Figure 3b presents the predicted distributions of six heavy metals obtained from the optimal

GWR-XGBoost model; $R^2$ ranged between 0.856 and 0.915, demonstrating the robustness of the predictions.

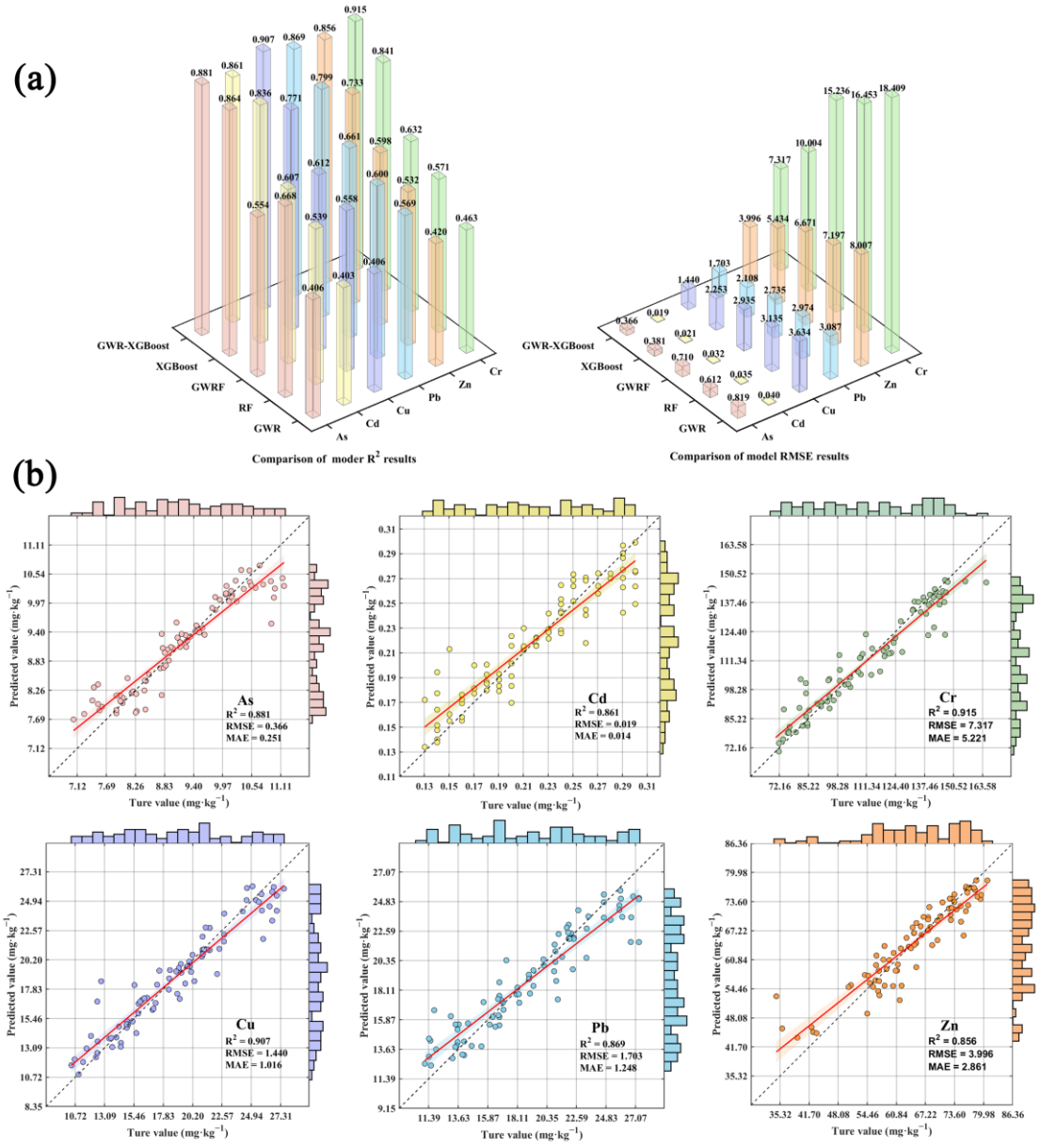



**Figure 3.** (**a**) $R^2$ and RMSE of six models for predicting soil heavy metal concentrations. (**b**) Assessment of heavy metal distribution patterns using the combined GWR-XGBoost model.

### *3.3. Environmental Drivers of Heavy Metal Contamination*

#### 3.3.1. Correlation Analysis

Pearson correlation was conducted to investigate the associations of soil heavy metals with various environmental covariates, providing initial insight into linear associations. TP was positively correlated with Cd, while NDVI exhibited negative associations with As and Cu. POP exhibited a positive correlation with Cr, while industrial activity—represented by CP and ALL—also showed positive correlations with Cr and Cu. These results indicate that local soils have been substantially affected by anthropogenic activities. To further explore the interrelationships among input features, correlation analyses were performed (Figure S9). Significant positive correlations were observed between AP

and ARH and Sand, RA, and POP, while NDVI and AT exhibited notable negative correlations with GTF and ETF.

### 3.3.2. SHAP Analysis

The SHAP method was employed to quantify the contribution of each environmental variable to soil heavy metal predictions. The feature importance analysis for Cd is presented in Figure 4. The corresponding results for the remaining heavy metals—As, Cu, Pb, Zn, and Cr—are shown in Figures S10–S14. The effects of major variables on the spatial distribution of heavy metals were further analysed (Text S2; Figure S15). The contributions of six heavy metals were analyzed. Soil physicochemical properties, such as TP and BD, indirectly reflect ecological conditions by representing fundamental soil characteristics [40]. Topographic factors, including slope and aspect, can affect the migration, accumulation, and dispersal patterns of soil heavy metals [28]. In addition, spatial covariates (SRIDW, SRTR, and SROK) made significant contributions to the prediction, effectively compensating for spatial heterogeneity during the modelling process. Meteorological covariates—including AWS, AP, ARH, and EVA—affect soil migration and dispersion processes, thereby indirectly characterising the environmental behaviour of soils [41]. Industrial activity covariates, such as GWT and ALL, also play significant roles. Consequently, the observed spatial variability of soil metals results from both natural pedogenic processes, including weathering and leaching [42], and human-induced factors such as industrial emissions, farming practices, transport, and urban development [43,44].

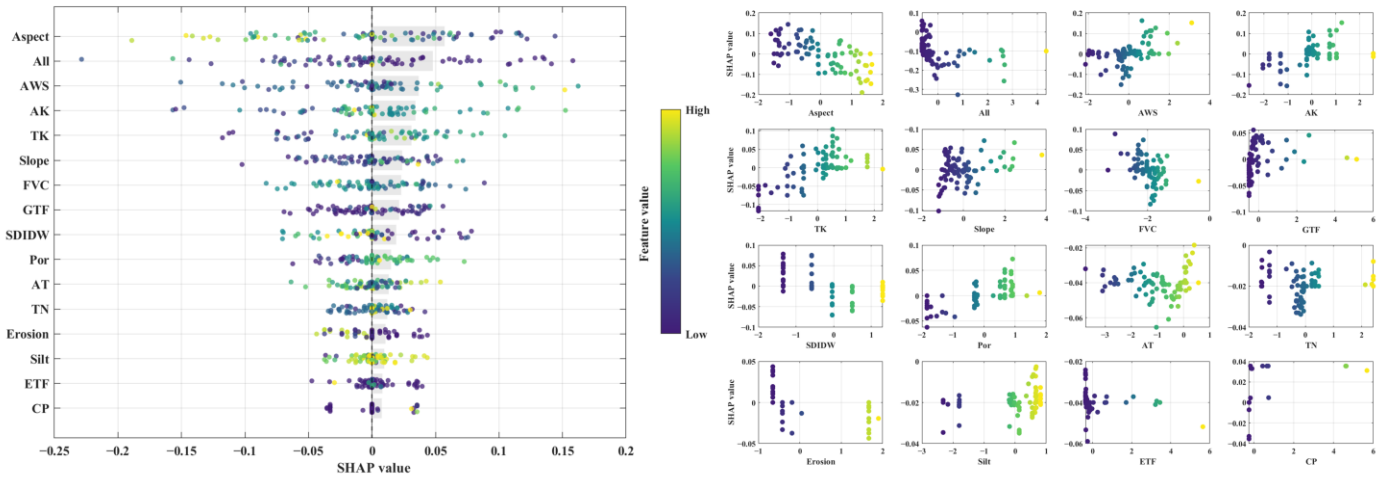



**Figure 4.** Importance analysis of input characteristics for heavy metal Cd.

### 3.3.3. Interactions Among Environmental Variables

SHAP interaction plots for six heavy metals reveal substantial differences in feature interactions among them (Figure 5). For Cu and Pb, interactions are primarily governed by the synergy and antagonism of topographic and soil properties, including slope, CEC, and FVC, although industrial activities (GTF) still exert some influence. This indicates that natural processes predominantly drive these metals while being modulated by anthropogenic disturbances [45,46]. For As and Cr, interactions are mainly associated with both natural and industrial influences, including AWS, ETF, and silt content. In areas of frequent industrial activity, acidic soils enhance the accumulation of As and Cr, highlighting the primary role of human activities [47,48]. The interactions of Cd and Zn are more complex, exhibiting strong interplay between soil properties and industrial activities, as well as close associations between transportation factors and natural conditions, reflecting the

influence of multiple sources on their spatial distributions [49,50]. SHAP interaction analysis was employed to quantify the interactions among environmental covariates, thereby effectively capturing the spatial heterogeneity of heavy metals and variations in their driving mechanisms. The findings help identify areas where heavy metal enrichment may be associated with specific environmental conditions or anthropogenic activities. These insights can inform targeted soil monitoring and help prioritize areas for pollution control and management [16,51].

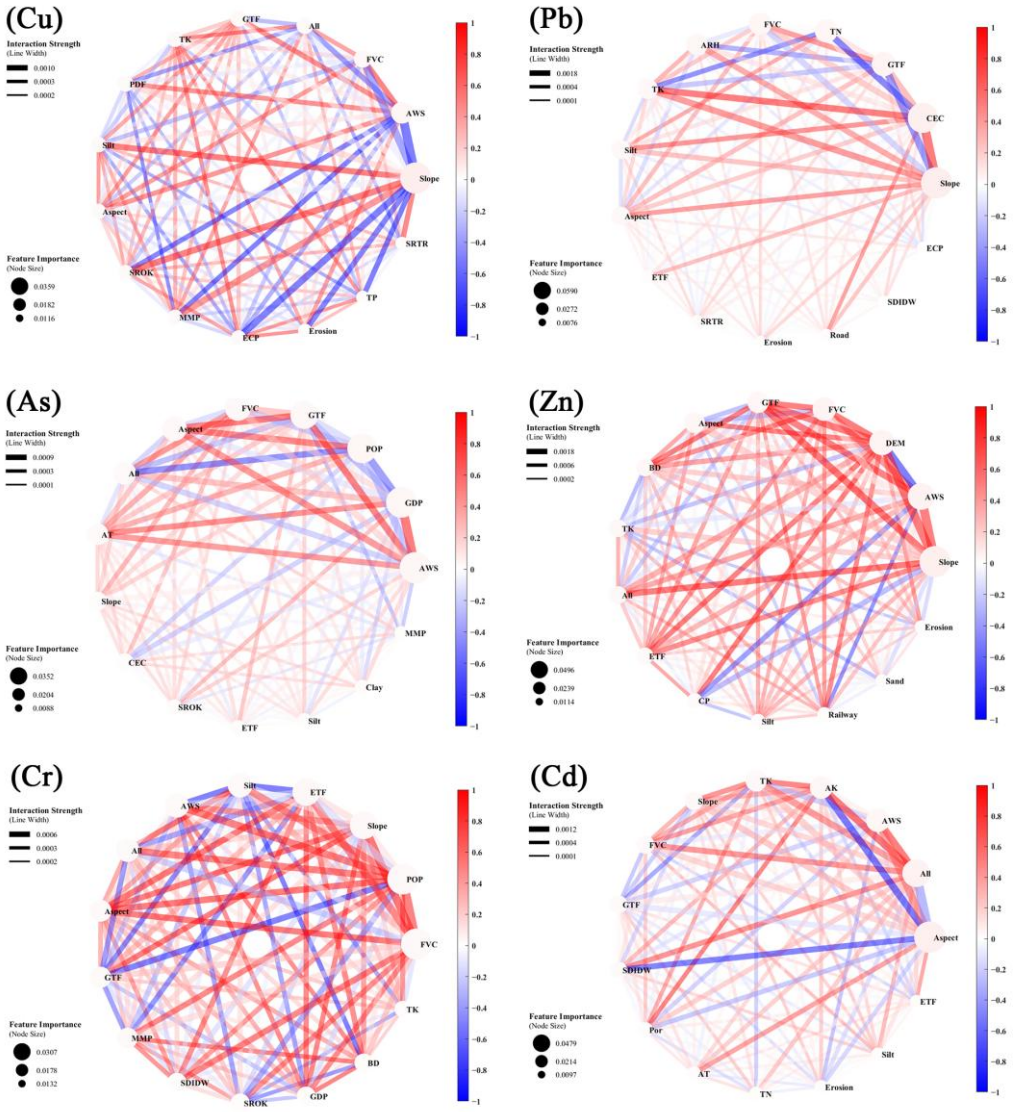



**Figure 5.** SHAP interaction effect analysis of heavy metals.

### *3.4. Spatial Prediction Results of Soil Heavy Metals*

Field sampling data, environmental variables, and multi-source spatial information were incorporated into a GWR-XGBoost framework, allowing for regional estimation of soil metal levels. Spatial distribution patterns of six heavy metals are presented in Figure 6. Except for Zn, the other five heavy metals exhibited pronounced pollution clustering in the Tiexi region. Gongzhuling displayed notable accumulation of As and Cr, whereas Yitong was characterized by Cd and Zn contamination. Severe contamination of Cr, Cu, and Pb was observed in Liaoyuan. High-concentration areas were predominantly distributed in urban regions, likely associated with urbanization and anthropogenic activities, including industrial emissions and agricultural production [52]. Furthermore, the spatial

distribution patterns of multiple heavy metals were strongly correlated with local population density (Figure S16).

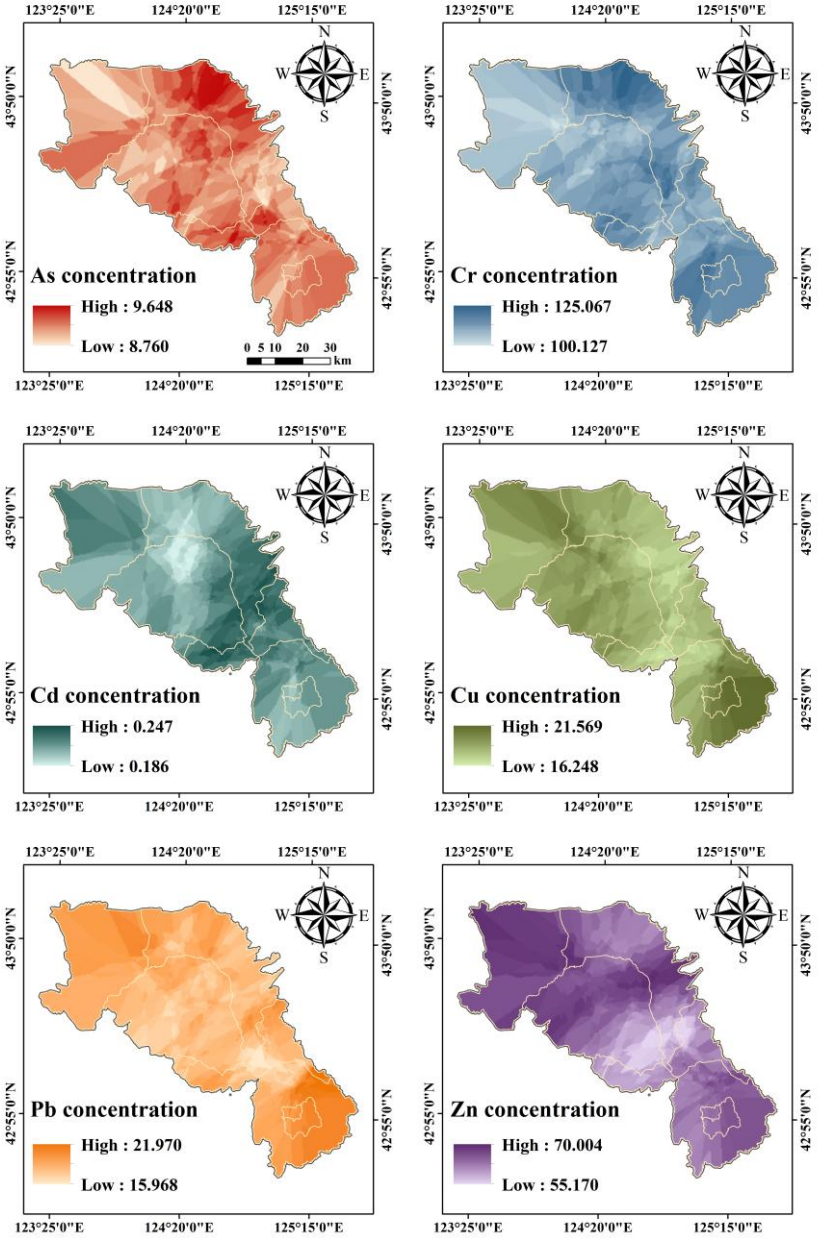



**Figure 6.** Predicted spatial distribution of various heavy metals using the optimal model.

### *3.5. Ecological Risk Assessment*

The predicted spatial distributions of soil heavy metals were used to assess contamination levels and potential ecological risks (Figure 7). Marked spatial heterogeneity was observed among different heavy metals. High risks of As, Cd, and Zn were mainly identified in eastern Gongzhuling, southwestern Shuangliao, and Tiexi District. Cr contamination was concentrated in eastern Gongzhuling and southern Longshan District. Cu-related risks were primarily distributed in northeastern Longshan District and southern Lishu County. Pb contamination was mainly concentrated in northwestern Shuangliao and southern Longshan District. The potential ecological risk factor of Cd in the study area ranged from 34.766 to 81.613. Approximately 82% of the sampling sites were classified as having moderate ecological risk, while 7% were categorized as having considerable ecological risk. In contrast, all other heavy metals were associated with low ecological risk

levels. The integrated ecological risk remained at a low level. This risk level was comparable to that reported for northeastern China [49], but lower than those observed in the northern Qinling Mountains, China [53] and East Java, Indonesia [54].

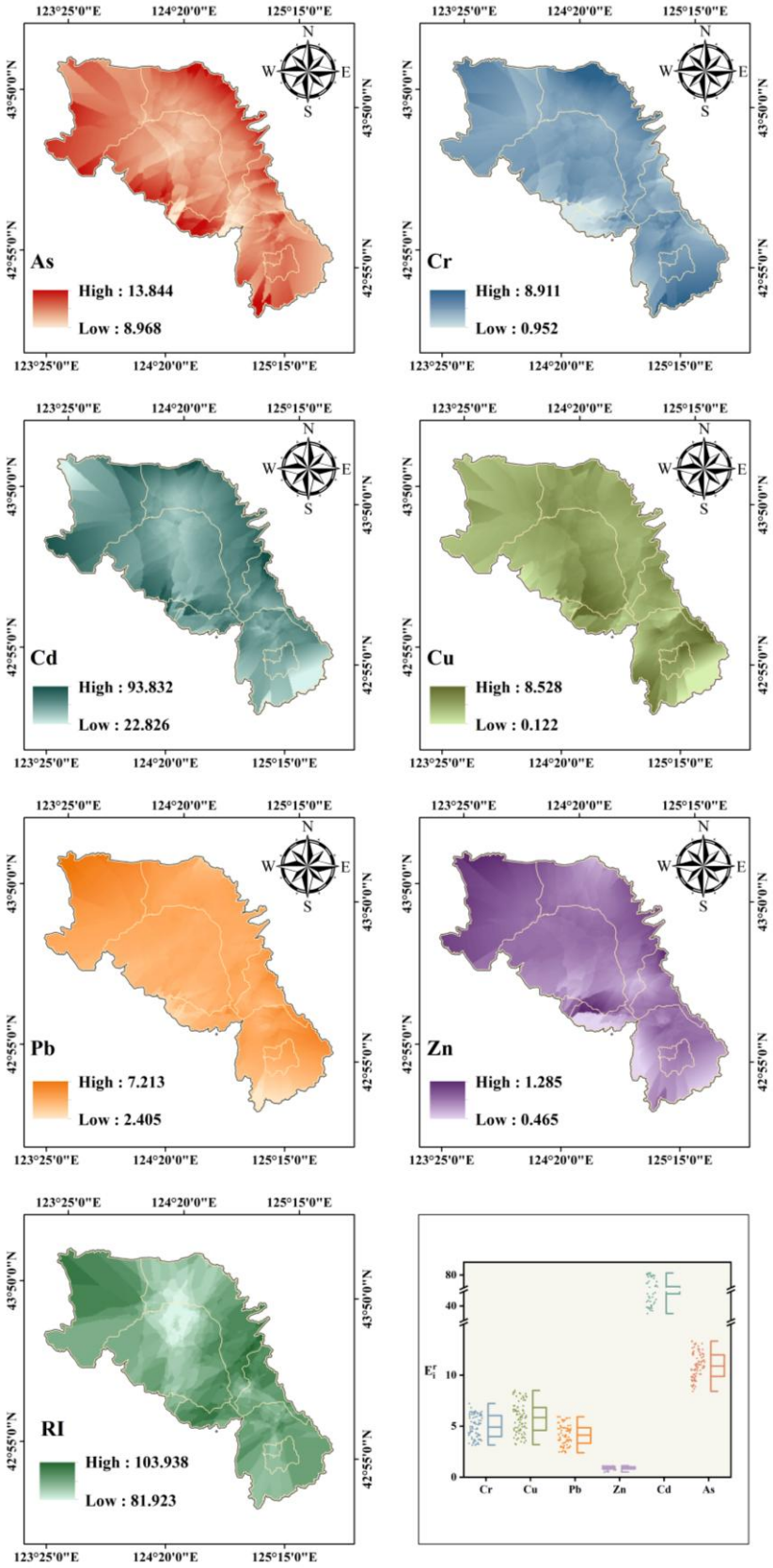



**Figure 7.** Results of the potential ecological risk index assessment.

### *3.6. Human Health Risk Assessment*

Monte Carlo simulations were conducted to characterize the probability distributions of carcinogenic and non-carcinogenic risks for adults and children (Figure 8). The carcinogenic risk values ranged from $1 \times 10^{-6}$ to $1 \times 10^{-4}$, indicating an acceptable risk level in the study area. The carcinogenic risk for children ($4.91 \times 10^{-5}$) was significantly higher than that for adults ($2.59 \times 10^{-5}$). The cumulative probability distributions of carcinogenic risks associated with As and Cr are shown in Figure 8c,d. Cr and As contributed substantially to the total carcinogenic risk, accounting for 40% and 32.5%, respectively (Figure 8e,f). Overall, the health risks did not exceed the thresholds established by the USEPA. Furthermore, the estimated health risks were substantially lower than those reported for Xiangfu District, Kaifeng, China [55] and Svirsk, Russia [47]. The study area as a whole remained within the acceptable carcinogenic risk range.

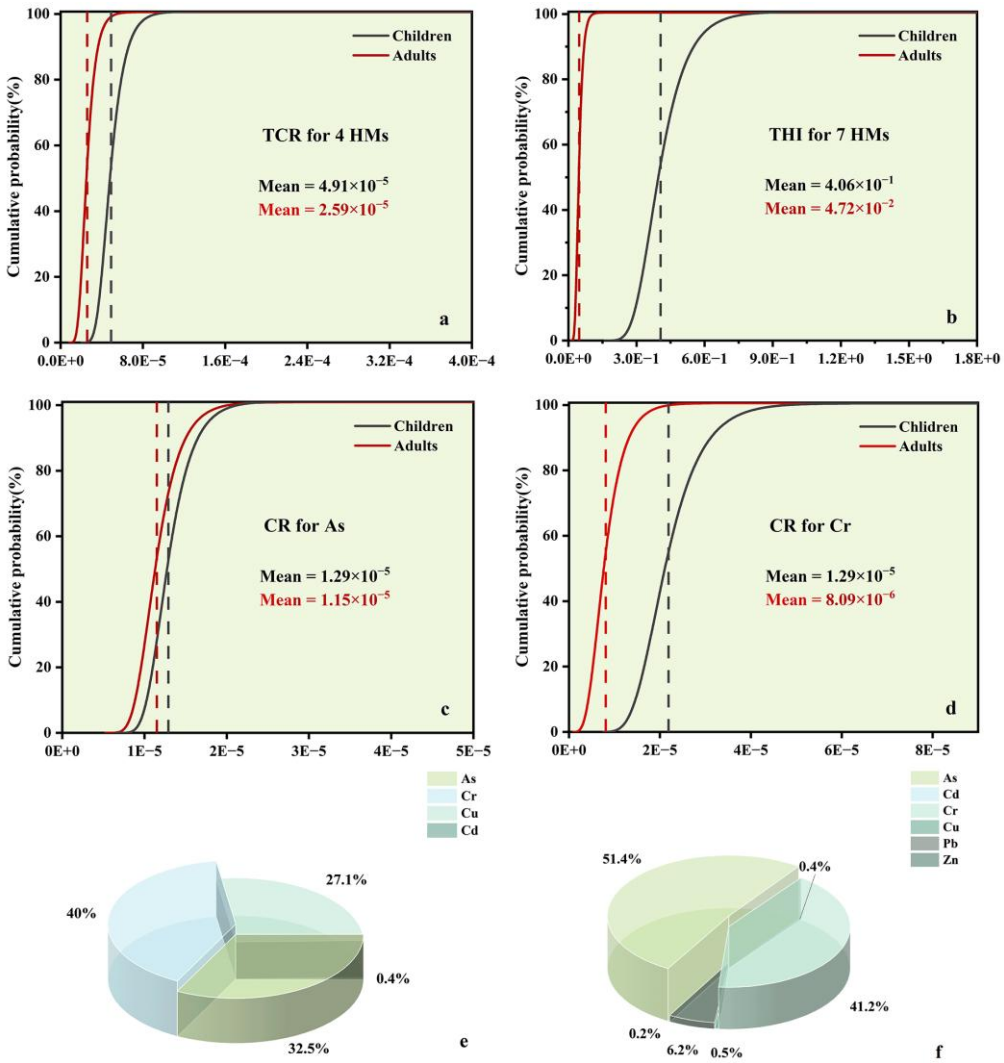



**Figure 8.** Cumulative probability distributions of health risks based on Monte Carlo simulation. (**a**) Cumulative probability distributions of carcinogenic risks for adults and children. (**b**) Cumulative probability distributions of non-carcinogenic risks for adults and children. (**c**) Cumulative probability distributions of carcinogenic risk associated with As for adults and children. (**d**) Cumulative probability distributions of carcinogenic risk associated with Cr for adults and children. (**e**) Proportions of four heavy metals contributing to the total carcinogenic risk. (**f**) Proportions of six heavy metals contributing to the total non-carcinogenic risk.

### *3.7. Hotspot Identification and Management Recommendations*

Based on the total hazard index (THI), total carcinogenic risk (TCR), and potential ecological risk index (RI) risk indices combined with a bivariate Moran's I analysis, pollution hotspots were identified in Gongzhuling City, northern Yitong, Tiexi District, Tiedong District, and Dongliao County (Figure 9). Priority management should be implemented in high-risk areas to facilitate timely pollution control and prevention.

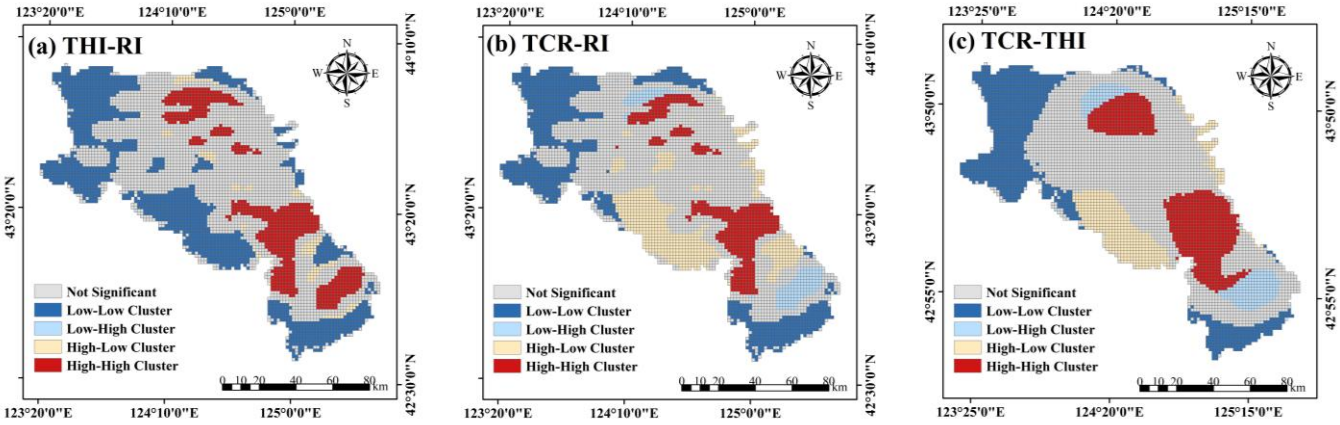



**Figure 9.** Bivariate local Moran's I. (**a**) THI-RI; (**b**) TCR-RI; (**c**) TCR-THI.

Targeted management measures should be implemented in the identified hotspot areas. The distributions of Cu and Zn were primarily influenced by natural factors, including soil properties and climatic conditions. In areas dominated by natural drivers, greater emphasis should be placed on soil and water conservation projects as well as soil quality management. In contrast, As and Cr were jointly affected by socioeconomic development and industrial activities. For areas where anthropogenic factors predominate, stricter industrial regulation and optimization of industrial structures are recommended. Such actions could reduce the risk of continuous heavy metal input into the environment. In hotspot regions such as Gongzhuling City and northern Yitong, enhanced health risk assessment and environmental monitoring should be conducted around densely populated areas and other sensitive receptors. Potential mitigation measures include controlling industrial emissions and wastewater discharges, as well as avoiding excessive fertilizer and pesticide use. In high-risk areas, appropriate soil amendments and low-accumulating crop varieties may help reduce heavy metal bioavailability and transfer to maize grains. Continued monitoring of heavy metal concentrations in soil and maize grains can help assess the effectiveness of these measures and track changes in potential exposure risks. Prioritizing these measures in identified hotspots and high-risk areas may help limit further contamination and protect human health.

### *3.8. Uncertainty Analysis and Research Prospects*

Several limitations remain in this study. (1) During spatial data processing and interpolation analyses, smoothing effects may occur, leading to attenuation of extreme values. Future studies could explore alternative spatial strategies to improve the representation of extreme conditions. (2) The selected multi-source variables remain limited. Incorporating additional datasets, such as remote sensing data, may further improve model interpretability and predictive accuracy.

## 4. Conclusions

This study integrated field observations, multi-source datasets, and machine learning approaches to identify the environmental drivers and spatial hotspots of soil heavy metal contamination. Following spatial block cross-validation and parameter optimization, the ensemble model demonstrated robust and reliable predictive performance. The results indicated that climatic variables, such as AWS, and soil properties, including TK and TN, were among the primary factors influencing heavy metal distributions. Moreover, significant interaction effects were observed among topographic, socioeconomic, and ecological variables, suggesting that soil heavy metal variability was jointly governed by multiple environmental drivers. Risk assessment revealed that the potential ecological risks associated with As, Cr, and Cd were widely distributed across the study area, with As and Cr identified as the dominant contributors to carcinogenic risks. Meanwhile, regions such as Gongzhuling City and northern Yitong County were identified as contamination hotspots requiring stringent management measures.

**Supplementary Materials:** The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/agriculture16212416/s1. Text S1: Method for obtaining the spatial covariance SRs metric; Text S2: Influence of key indicators on the spatial distribution of soil heavy metals;. Figure S1: Research framework of this study; Figure S2: Spatial distribution of soil properties; Figure S3: Spatial distribution of topographic conditions; Figure S4: Spatial distribution of climate conditions; Figure S5: Spatial distribution of ecological functions; Figure S6: Spatial distribution of social economy; Figure S7: Spatial distribution of industrial activity; Figure S8: Spatial patterns of soil heavy metal levels; Figure S9: Correlation analysis of input characteristics for soil heavy metal prediction; Figure S10: Importance analysis of input characteristics for heavy metal As; Figure S11: Importance analysis of input characteristics for heavy metal Cu; Figure S12: Importance analysis of input characteristics for heavy metal Pb; Figure S13: Importance analysis of input characteristics for heavy metal Zn; Figure S14: Importance analysis of input characteristics for heavy metal Cr; Figure S15: Schematic diagram of spatial distribution for the top eight key variables; Figure S16: Spatial distribution map of population density in the Liaohe River Basin, Jilin Province; Table S1: Details of variables affecting heavy metal contamination of soil; Table S2: Correlation analysis between soil heavy metals and influencing factors.

**Author Contributions:** Conceptualization, Y.C., F.Z. and J.Z.; methodology, C.Z., J.Z. and H.L.; software, H.L. and G.A.; validation, G.A.; formal analysis, J.W.; data curation, J.W.; writing—original draft preparation, J.W., C.Z.; writing—review and editing, C.Z.; visualization, J.W.; supervision, Y.C. and F.Z.; funding acquisition, C.Z., F.Z. and J.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Natural Science Foundation of China Joint Fund Project (No. U21A2040), the National Natural Science Foundation of China (No.42077443), the major science and technology program of Jilin Province (No. 20230303009SF) and the Science and Technology Research Project of the Education Department of Jilin Province (No. JJKH20261774KJ).

**Institutional Review Board Statement:** Not applicable.

**Data Availability Statement:** The raw data supporting the conclusions of this article will be made available by the authors on request.

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

## References

1. FAO. *Healthy Soils Are the Basis for Healthy Food Production*; Food and Agriculture Organization of the United Nations: Rome, Italy, 2015.

2. Hou, D.; Jia, X.; Wang, L.; McGrath, S.P.; Zhu, Y.; Hu, Q.; Zhao, F.; Bank, M.S.; Connor, D.O.; Nriagu, J. Global soil pollution by toxic metals threatens agriculture and human health. *Science* **2025**, *388*, 316–321. https://doi.org/10.1126/science.adr5214.
3. Miao, Y.; Li, H.; Xue, L.; Shen, C.; Wang, F. Application of a semi-variogram-based KNN algorithm in the spatial prediction of soil heavy metals. *Environ. Pollut.* **2026**, *390*, 127436. https://doi.org/10.1016/j.envpol.2025.127436.
4. Prabagar, S.; Dharmadasa, R.M.; Lintha, A.; Thuraisingam, S.; Prabagar, J. Accumulation of heavy metals in grape fruit, leaves, soil and water: A study of influential factors and evaluating ecological risks in Jaffna, Sri Lanka. *Environ. Sustain. Indic.* **2021**, *12*, 100147. https://doi.org/10.1016/j.indic.2021.100147.
5. Tarighat, H.; Baghdadi, M.; Nasrabadi, T. Integrating native reference values with fractionation and multi-index analysis for heavy metal risk assessment in industrial soils of the South Pars Economic Energy Zone, Iran. *Environ. Pollut.* **2026**, *390*, 127543. https://doi.org/10.1016/j.envpol.2025.127543.
6. Zhou, B.; Wang, F.; Li, H.; Zhao, Y.; Yang, R.; Huang, H.; Wang, Y.; Xiao, Z.; Tian, K.; Pang, W. Evaluating heavy metals-related risk in staple crops and making financing strategy for corresponding soil remediation across China. *J. Hazard. Mater.* **2024**, *480*, 136135. https://doi.org/10.1016/j.jhazmat.2024.136135.
7. Lu, X.; Chen, X.; Vancov, T.; Zhu, F.; Zhu, W.; Hong, L.; Yao, Y.; Li, P.; Wang, W.; Hong, C. Combined remediation effect of ryegrass-earthworm on heavy metal composite contaminated soil. *J. Hazard. Mater.* **2025**, *494*, 138477. https://doi.org/10.1016/j.jhazmat.2025.138477.
8. Xu, Y.; Li, P.; Zhang, Z.; Gu, Y.; Xiao, L.; Liu, X.; Wang, B. Integrating machine learning for enhanced spatial prediction and risk assessment of soil heavy metal(loid)s. *Environ. Pollut.* **2025**, *383*, 126919. https://doi.org/10.1016/j.envpol.2025.126919.
9. Zeng, L.; Wang, Y.; Jing, L.; Cheng, Q. Quantitative determination of auxiliary information for mapping soil heavy metals and soil contamination risk assessment. *Appl. Geochem.* **2021**, *130*, 104964. https://doi.org/10.1016/j.apgeochem.2021.104964.
10. Wang, H.; Yilihamu, Q.; Yuan, M.; Bai, H.; Xu, H.; Wu, J. Prediction models of soil heavy metal(loid)s concentration for agricultural land in Dongli: A comparison of regression and random forest. *Ecol. Indic.* **2020**, *119*, 106801. https://doi.org/10.1016/j.ecolind.2020.106801.
11. Zhang, H.; Yin, A.; Yang, X.; Fan, M.; Shao, S.; Wu, J.; Wu, P.; Zhang, M.; Gao, C. Use of machine-learning and receptor models for prediction and source apportionment of heavy metals in coastal reclaimed soils. *Ecol. Indic.* **2021**, *122*, 107233. https://doi.org/10.1016/j.ecolind.2020.107233.
12. Fan, X.; Gao, P.; Wang, J.; Zhang, M.; Hong, S.; Qin, S.; Zhang, Q.; Ma, L.; Zhang, L.; Zhang, Z. et al. Assessment and prediction of soil heavy metal pollution in cotton fields by multi-source data feature fusion combined with machine learning. *Ind. Crop. Prod.* **2025**, *233*, 121447. https://doi.org/10.1016/j.indcrop.2025.121447.
13. Han, Z.; Wang, J.; Liao, X.; Yang, J. Accurate prediction of spatial distribution of soil heavy metal in complex mining terrain using an improved machine learning method. *J. Hazard. Mater.* **2025**, *491*, 137994. https://doi.org/10.1016/j.jhazmat.2025.137994.
14. Zantvoort, K.; Nacke, B.; Görlich, D.; Hornstein, S.; Jacobi, C.; Funk, B. Estimation of minimal data sets sizes for machine learning predictions in digital mental health interventions. *npj Digit. Med.* **2024**, *7*, 361. https://doi.org/10.1038/s41746-024-01360-w.
15. Padarian, J.; McBratney, A.B.; Minasny, B. Game theory interpretation of digital soil mapping convolutional neural networks. *SOIL* **2020**, *6*, 389–397. https://doi.org/10.5194/soil-6-389-2020.
16. Yan, Y.; Yang, Y. Revealing the synergistic spatial effects in soil heavy metal pollution with explainable machine learning models. *J. Hazard. Mater.* **2025**, *482*, 136578. https://doi.org/10.1016/j.jhazmat.2024.136578.
17. Pan, Y.; Sha, A.; Han, W.; Liu, C.; Liu, G.; Welsch, E.; Zeng, M.; Xu, S.; Zhao, Y.; Tian, S. et al. Identifying spatial drivers of soil heavy metal pollution risk integrating positive matrix factorization, machine learning, and multi-scale geographically weighted regression. *J. Hazard. Mater.* **2025**, *485*, 136841. https://doi.org/10.1016/j.jhazmat.2024.136841.
18. Shangguan, W.; 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, https://doi.org/10.1002/jame.20026.
19. Kummu, M.; Kosonen, M.; Masoumzadeh Sayyar, S. Downscaled gridded global dataset for gross domestic product (GDP) per capita PPP over 1990–2022. *Sci. Data* **2025**, *12*, 178. https://doi.org/10.1038/s41597-025-04487-x.
20. Zhao, W.; Ma, J.; Liu, Q.; Dou, L.; Qu, Y.; Shi, H.; Sun, Y.; Chen, H.; Tian, Y.; Wu, F. Accurate Prediction of Soil Heavy Metal Pollution Using an Improved Machine Learning Method: A Case Study in the Pearl River Delta, China. *Environ. Sci. Technol.* **2023**, *57*, 17751–17761. https://doi.org/10.1021/acs.est.2c07561.
21. Yang, B.; He, A.; Ren, Z.; Yu, K.; Zhao, G.; Fan, Y.; Wang, Q.; Luo, S. A transfer learning-enhanced deep learning framework for efficient and interpretable soil heavy metal pollution prediction under data scarcity and spatial heterogeneity. *J. Hazard. Mater.* **2025**, *495*, 138926. https://doi.org/10.1016/j.jhazmat.2025.138926.

22. Li, J.; Heap, A.D.; Potter, A.; Daniell, J.J. Application of machine learning methods to spatial interpolation of environmental variables. *Environ. Model. Softw.* **2011**, *26*, 1647–1659. https://doi.org/10.1016/j.envsoft.2011.07.004.
23. 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. https://doi.org/10.1016/j.jhazmat.2024.133745.
24. Fotheringham, A.S.; Charlton, M.; Brunsdon, C. The geography of parameter space: An investigation of spatial non-stationarity. *Geogr. Inf. Syst.* **1996**, *10*, 605–627. https://doi.org/10.1080/026937996137909.
25. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. https://doi.org/10.1023/A:1010933404324.
26. Qin, Z.; Peng, Q.; Jin, C.; Xu, J.; Xing, S.; Zhu, P.; Yang, G. Geographically weighted random forest fusing multi-source environmental covariates for spatial prediction of soil heavy metals. *Environ. Pollut.* **2025**, *385*, 127135. https://doi.org/10.1016/j.envpol.2025.127135.
27. Chen, T.; Guestrin, C. *XGBoost: A Scalable Tree Boosting System*; ACM: New York, NY, USA, 2016, pp. 785–794. https://doi.org/10.1145/2939672.2939785.
28. Li, J.; Li, X.; Wu, J.; Wang, C.; Liu, J.; Zang, F. Precision mapping and driving factors of heavy metal(loid)s in agricultural soils of the Yellow River: An integrated machine learning and Geodetector approach. *Environ. Res.* **2026**, *292*, 123681. https://doi.org/10.1016/j.envres.2026.123681.
29. Georganos, S.; Grippa, T.; Gadiaga, A.N.; Linard, C.; Lennert, M.; Vanhuysse, S.; Mboga, N.; Eléonore Wolff, E.; Kalogirou, S. Geographical random forests: A spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto Int.* **2019**, *36*, 121–136. https://doi.org/10.1080/10106049.2019.1595177.
30. Brunsdon, C.; Fotheringham, A.S.; Charlton, M.E. Geographically Weighted Regression A Method for Exploring Spatial Non-stationarity. *Geogr. Anal.* **1996**, *28*, 282–298. https://doi.org/10.1111/j.1538-4632.1996.tb00936.x.
31. Lundberg, S.M.; Lee, S. A Unified Approach to Interpreting Model Predictions. In *31st Conference on Neural Information Processing Systems*; ACM: New York, NY, USA, 2017.
32. Jilin Environmental Monitoring Center; Jilin Environmental Protection Research Institute; Changchun Geographic Research Institute, Chinese Academy of Sciences. *Study on the Soil Background Values of Jilin Province*; EMCS: Changchun, China, 1990.
33. Hakanson, L. An ecological risk index for aquatic pollution control. A sedimentological approach. *Water Res.* **1980**, *14*, 975–1001.
34. Kamunda, C.; Mathuthu, M.; Madhuku, M. Health Risk Assessment of Heavy Metals in Soils from Witwatersrand Gold Mining Basin, South Africa. *Int. J. Environ. Res. Public Health* **2016**, *13*, 663. https://doi.org/10.3390/ijerph13070663.
35. Sun, P.; Chen, Y.; Wang, X.; Zhou, Z.; Zhu, X.; Sun, S.; Xu, J. Quantification of an integrated approach to heavy metal source apportionment and probabilistic health risk assessment in the black soil region of central Jilin Province, China. *Ecotoxicol. Environ. Saf.* **2025**, *299*, 118358. https://doi.org/10.1016/j.ecoenv.2025.118358.
36. USEPA Exposure Factors. *Handbook: 2011Edition: Offce of Research and Development*; USEPA: Washington, DC, USA, 2011.
37. Chen, H.; Teng, Y.; Lu, S.; Wang, Y.; Wu, J.; Wang, J. Source apportionment and health risk assessment of trace metals in surface soils of Beijing metropolitan, China. *Chemosphere* **2016**, *144*, 1002–1011. https://doi.org/10.1016/j.chemosphere.2015.09.081.
38. Yuan, B.; Cao, H.; Du, P.; Ren, J.; Chen, J.; Zhang, H.; Zhang, Y.; Luo, H. Source-oriented probabilistic health risk assessment of soil potentially toxic elements in a typical mining city. *J. Hazard. Mater.* **2023**, *443*, 130222. https://doi.org/10.1016/j.jhazmat.2022.130222.
39. Luo, H.; Yang, L.; Zhang, C.; Xiao, X.; Lyu, X. Early warning of heavy metals contamination in agricultural soils: Spatio-temporal distribution and future trends in the Hexi Corridor. *Ecol. Indic.* **2024**, *160*, 111908. https://doi.org/10.1016/j.ecolind.2024.111908.
40. Liu, H.; Xiong, Z.; Jiang, X.; Liu, G.; Liu, W. Heavy metal concentrations in riparian soils along the Han River, China: The importance of soil properties, topography and upland land use. *Ecol. Eng.* **2016**, *97*, 545–552. https://doi.org/10.1016/j.ecoleng.2016.10.060.
41. Zhang, L.; Wang, B.; Wu, W.; Wang, C.; Cheng, H.; Duan, X. Enhanced health risk of soil heavy metal exposure following an extreme rainstorm under climate change. *Sci. Total Environ.* **2024**, *954*, 176409. https://doi.org/10.1016/j.scitotenv.2024.176409.
42. Risueño, Y.; Petri, C.; Conesa, H.M. The importance of edaphic niches functionality for the sustainability of phytomanagement in semiarid mining impacted ecosystems. *J. Environ. Manag.* **2020**, *266*, 110613. https://doi.org/10.1016/j.jenvman.2020.110613.
43. Li, Y.; Zhang, J.; Song, N.; Wang, Y.; Yu, J.; He, L.; Yang, R.; Yang, L.; He, D. Assessment of health risk and identification of pollution sources of heavy metals in water in Chongqing's wastewater treatment plants based on ICP-MS. *Environ. Pollut.* **2025**, *373*, 126193. https://doi.org/10.1016/j.envpol.2025.126193.

44. Xu, Z.; Yin, M.; Yang, X.; Yang, Y.; Xu, X.; Li, H.; Hong, M.; Qiu, G.; Feng, X.; Tan, W. et al. Simulation of vertical migration behaviors of heavy metals in polluted soils from arid regions in northern China under extreme weather. *Sci. Total Environ.* **2024**, *919*, 170494. https://doi.org/10.1016/j.scitotenv.2024.170494.
45. Liu, N.; Li, X.; Chen, P.; Yuan, W.; Wang, D.; Wang, X. Climate and vegetation controlling accumulation and translocation of heavy metals in water tower regions of Qinghai-Tibet Plateau. *J. Hazard. Mater.* **2025**, *484*, 136752. https://doi.org/10.1016/j.jhazmat.2024.136752.
46. Tong, B.; Li, Y.; Zhu, G.; Han, X.; Li, J.; Yan, Y.; Chen, Y.; Chen, H.; Zhai, Y. Integration of Geodetector and PMF model improves source identification of regional soil heavy metals. *J. Hazard. Mater. Adv.* **2026**, *23*, 101283. https://doi.org/10.1016/j.hazadv.2026.101283.
47. Baenguev, B.; Belogolova, G.; Chuparina, E.; Minkina, T.; Kravchenko, E. Arsenic and heavy metals in reclaimed soils of sulfide-containing waste dumps: Contamination levels, spatial distribution, bioavailability, and health risk assessment. *Environ. Res.* **2026**, *301*, 124544. https://doi.org/10.1016/j.envres.2026.124544.
48. Zhou, Y.; Ding, D.; Li, X.; Fan, T.; Liu, M.; Deng, S.; Zhang, S.; Xia, F.; Kong, L. The integrating eco-health risk assessment and driving factors for risks of heavy metals in soils from an open-pit lead-zinc mine area. *Ecotox. Environ. Saf.* **2026**, *321*, 120348. https://doi.org/10.1016/j.ecoenv.2026.120348.
49. Che, T.; Deng, B.; Hu, N.; Wu, M.; Yu, H.; Yue, J.; Wang, Q. Spatiotemporal evolution and ecological risk assessment of heavy metals in agricultural black soils of Northeast China. *Environ. Res.* **2026**, *301*, 124567. https://doi.org/10.1016/j.envres.2026.124567.
50. Shi, R.; Qiang, Y.; Chu, C.; Li, B.; Xiao, Y.; Zhang, N.; Guan, Y. Assessing the heavy metal(loid) pollution risk potential in soils: A spatially explicit and multi-source integrated framework. *Environ. Impact Assess. Rev.* **2026**, *121*, 108584. https://doi.org/10.1016/j.eiar.2026.108584.
51. Li, W.; Jiang, Y.; Ye, Y.; Guo, X.; Shi, Z. Spatiotemporal interpretable mapping framework for soil heavy metals. *J. Clean. Prod.* **2024**, *468*, 143101. https://doi.org/10.1016/j.jclepro.2024.143101.
52. Shen, C.; Huang, S.; Wang, M.; Wu, J.; Su, J.; Lin, K.; Chen, X.; He, T.; Li, Y.; Sha, C. et al. Source-oriented health risk assessment and priority control factor analysis of heavy metals in urban soil of Shanghai. *J. Hazard. Mater.* **2024**, *480*, 135859. https://doi.org/10.1016/j.jhazmat.2024.135859.
53. Huo, B.; Shen, Y.; Ning, Y.; Shen, W.; Ma, P.; Jia, Z.; Zhao, L.; Zhao, J.; Chen, X.; Ding, T.; et al. Accumulation pattern, source apportionment and risk assessment of soil foothills of Qinling mountains, China. *CATENA* **2026**, *266*, 109938. https://doi.org/10.1016/j.catena.2026.109938.
54. Purnama, D.S.; Nur'Azizah Rahman, I.; Santoso, M.; Syafei, A.D.; Lestiani, D.D.; Kurniawati, S.; Kusmartini, I.; Nurhaini, F.F.; Syahfitri, W.Y.N.; Atmodjo, D.P.D. et al. Spatial distributions and risk assessments of heavy metals and naturally occurring radionuclides in soils of East Java, Indonesia. *Environ. Chem. Ecotoxicol.* **2026**, *8*, 1243–1255. https://doi.org/10.1016/j.enceco.2026.01.022.
55. Leng, W.; Li, L.; Liu, H.; Bai, L.; Zhang, M.; Wang, L.; Liu, J. Source-specific risks of heavy metals in urban–agricultural soils and implications for human health in a liver cancer high-incidence area of Central China. *J. Clean. Prod.* **2026**, *554*, 148109. https://doi.org/10.1016/j.jclepro.2026.148109.


**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.