

Article

Spatial Distribution Prediction of Soil Heavy Metals Based on Random Forest Model

Shunqi Nie, Honghua Chen *, Xinxin Sun and Yunce An

School of Civil Engineering, Nanjing Forestry University, Nanjing 210037, China

* Correspondence: chenhh@njfu.edu.cn

Abstract: Mastering the spatial distribution of soil heavy metal content and evaluating the pollution status of soil heavy metals is of great significance for ensuring agricultural production and protecting human health. This study used a machine learning model to study the spatial distribution of soil heavy metal content in a coastal city in eastern China. Having obtained six soil heavy metal contents, including Cr, Cd, Pb, As, Hg, and Ni, environmental variables such as precipitation, soil moisture, and population density were selected. Random forest (RF) was used to model the spatial distribution of soil heavy metal content. The research findings indicate that the RF model demonstrates a robust predictive capability in discerning the spatial distribution of soil heavy metals, and environmental factor variables can explain 60%, 52.3%, 53.5%, 63.1%, 61.2%, and 51.2% of the heavy metal content of Cr, Cd, Pb, As, Hg, and Ni in soil, respectively. Among the chosen environmental variables, precipitation and population density exert notable influences on the predictive outcomes of the model. Specifically, precipitation exhibits the most substantial impact on Cr and Ni, whereas population density emerges as the primary determinant for Cd, Pb, As, and Hg. The RF prediction results show that Cr and Ni in the study area are less affected by human activities, while Cd, Pb, As, and Hg are more affected by human industrial and agricultural production. Research has shown that using RF models for predicting soil heavy metal distributions has certain significance.

Keywords: soil heavy metals; random forest; spatial interpolation; spatial distribution prediction



Citation: Nie, S.; Chen, H.; Sun, X.; An, Y. Spatial Distribution Prediction of Soil Heavy Metals Based on Random Forest Model. *Sustainability* **2024**, *16*, 4358. <https://doi.org/10.3390/su16114358>

Academic Editor: Georgios Koubouris

Received: 2 April 2024
Revised: 18 May 2024
Accepted: 20 May 2024
Published: 22 May 2024



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

1. Introduction

Soil plays a significant role in ecosystem balance [1], crop yield [2], and human health [3]. With the advancement of modern industry and agricultural activities such as irrigation and fertilization, soil heavy metal pollution has become increasingly severe [4,5]. Characterized by difficulties in degradation, a propensity for accumulation, and significant toxicity, soil heavy metal pollution has emerged as a pressing global environmental concern [6–8]. The accumulation of soil heavy metals poses a significant environmental challenge, resulting in various adverse impacts on ecosystems [9]. Excessive levels of soil heavy metals not only impede soil productivity but also pose risks to human and animal health through their inhalation and ingestion pathways [10,11]. Thus, understanding the spatial distribution of soil heavy metal contents and evaluating the status of soil heavy metal pollution are crucial for safeguarding agricultural production and protecting human health [12–15].

Currently, a considerable amount of research has been conducted on the spatial distribution prediction of soil heavy metal. Classic spatial interpolation methods which are used in soil heavy metal prediction include Kriging and Geographically Weighted Regression (GWR) [16]. In recent years, there has been increasing attention towards utilizing machine learning models, such as the k-nearest neighbors algorithm (KNN) [17], classification and regression tree models (CART) [18], tree models [19], and RF models [20], among others. These methods aid in better understanding and accurately predicting the spatial variability of soil characteristics. In comparison with various machine learning models, the RF

model demonstrates advantages due to its resistance to overfitting and insensitivity to multicollinearity [20]. For example, Dharumarajan [21] successfully employed an RF model to predict soil organic carbon, pH, and other soil attributes in southern India. Chagas [22] used an RF model to predict soil surface texture in a semiarid region and found that the RF model performed with higher accuracy than multiple linear regression. Guo [23] discovered their RF model performed better than stepwise linear regression and a generalized additive mixed model in predicting the soil total nitrogen of a rubber plantation. This indicates that an RF model is a more reliable and effective approach for the spatial prediction of soil characteristics. Although RF models have been widely used for the spatial prediction of soil characteristics, current studies have mostly focused on predicting soil heavy metal content in relatively small areas [23–25]. This underscores the need to investigate the accuracy of the RF model in predicting soil heavy metal content in large-scale studies.

Multiple environmental factors collectively contribute to the spatial distribution of soil heavy metals [26,27]. Existing research primarily selects environmental factors across both natural and socio-economic aspects. From a natural perspective, soil moisture [28], vegetation coverage [21], organic matter [29,30], elevation, and slope [31] are considered important factors influencing soil heavy metal content. In addition, precipitation also plays an important role in soil heavy metal content through leaching and adsorption processes [32–35]. As for socio-economic aspects, population density and the distance to transportation were chosen to predict the spatial characteristics of heavy metals in farmland soil [36]. Guo [37] indicated a negative correlation between heavy metals and the distance to roads within a certain range. Liu [38] demonstrated that population density reflects the intensity of living activities, suggesting that areas with a higher population density are likely to induce more heavy metal contamination.

Taking a coastal city in eastern China as the study area, this research utilizes Geographic Information System (GIS) and remote sensing (RS) techniques to extract environmental factor variables including elevation, slope, precipitation, soil moisture, organic matter, NDVI, distance to road, and population density. The field sampling of data is employed, with the content of six heavy metals—Chromium (Cr), Cadmium (Cd), Lead (Pb), Arsenic (As), Mercury (Hg), and Nickel (Ni)—serving as predictive targets. A correlation analysis is conducted to investigate the relationship between the environmental factor variables and the predictive targets. By employing random forest (RF) modeling and selecting optimal model parameters, a predictive model for soil heavy metals is constructed, enabling the prediction of their spatial distribution. The study results are expected to provide a data foundation for the assessment and prevention of soil heavy metal pollution.

2. Materials and Methods

2.1. Study Area

The study area is a coastal city in eastern China, which is located in a hilly region characterized by continuous hills and crisscrossing ravines, with higher terrain in the northeast, central, and western parts (Figure 1). The soil types in the study area are brown soil, skeletal soil, and cinnamon soil, with metamorphic rock as the underlying geology. It has a temperate continental climate, with an average annual temperature of 11.5 °C and an average annual rainfall of 671.1 mm. The average elevation is 136 m. The population density in the study area is 395 people per square kilometer, with a road length of approximately 2300 km. Planning one sampling point per 1 km × 1 km, we followed the principle of comprehensive distribution in terms of soil, land use type, and spatial coverage. This guaranteed that each type of soil and land use in the study area had surface sampling points, achieving complete spatial coverage and avoiding large unsampled areas within the region. Soil samples were collected from the surface layer, with a sampling depth of 0–20 cm. The northwestern part of the study area has sparse sampling points, mainly due to its mountainous terrain, sparse population, and predominant land use type being forested areas, where the heavy metal content in the soil is less influenced by environmental factors, hence resulting in fewer sampling points.

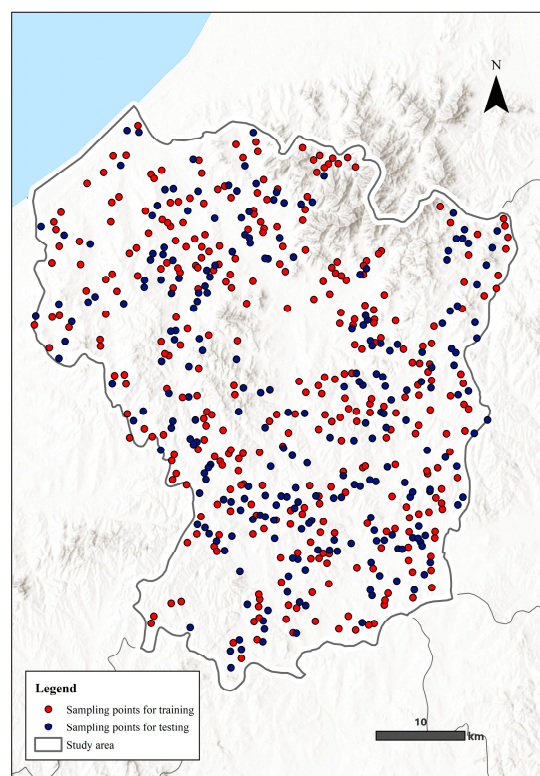


Figure 1. Study area and sample points.

2.2. Data

According to the “National Standard of the People’s Republic of China for Soil Environmental Quality Risk Control Standard for Agricultural Land Soil Pollution (Trial)” (GB 15618-2018) [39], Cr, Cd, Pb, As, Hg, and Ni have been selected as the six heavy metal elements for spatial interpolation.

In this study, soil heavy metal content data for Cr, Cd, Pb, As, Hg, and Ni were obtained using an Olympus handheld XRF soil heavy metal analyzer in the study area. After data cleaning, a total of 748 sampling points were obtained. Additionally, environmental factor variables for the study area were collected, including the soil organic matter, Digital Elevation Model (DEM), annual average precipitation [40], annual average soil humidity [41], slope, Normalized Difference Vegetation Index (NDVI) [42], land use type, population density [43], distance to roads, and distance to residential areas. The sources of these environmental variables are as follows: The DEM and slope data were obtained from the Geographic Spatial Data Cloud Platform, with a spatial resolution of 30 m. The annual average precipitation and annual average soil humidity data were obtained from the National Tibetan Plateau Data Center, with a spatial resolution of 1 km. NDVI data were obtained from the National Ecological Science Data Center, with a spatial resolution of 30 m. The population density data were obtained from the WorldPop dataset, with a spatial resolution of 1 km. Road data were obtained from OpenStreetMap. The distance to residential areas was calculated using ArcGIS proximity analysis with the “residential areas” category as the land use type, determining the distance from each sampling point to the nearest residential area. (See Table 1).

Table 1. List of environmental indicators.

Name of Indicators	Unit	Data Sources	Calculation Method
Elevation	m	www.gscloud.cn (accessed on 1 October 2023)	/
Slope	degree	www.gscloud.cn	/
Precipitation	mm	https://doi.org/10.5281/zenodo.3185722 (accessed on 1 October 2023)	/
Humidity	dm ³ /m ³	https://cstr.cn/18406.11.Terre.tpd.272415 (accessed on 1 October 2023)	/
Organic Matter	mg/kg	https://doi.org/10.11888/Soil.tpd.270281 (accessed on 1 October 2023)	/
NDVI	/	https://cstr.cn/15732.11.nesdc.ecodb.rs.2021.012 (accessed on 1 October 2023)	/
Distance to Living Area	m	/	The distance from sampling points to residential areas is calculated based on the land use type by a “proximity analysis” in ArcGIS 10.8.
Distance to Road	m	https://www.openstreetmap.org (accessed on 1 October 2023)	The distance from sampling points to roads is calculated by a “proximity analysis” in ArcGIS 10.8.
Population Density	person/km ²	www.worldpop.org (accessed on 1 October 2023)	/

2.3. The Random Forest Model Based on Environmental Factors

RF is an ensemble classification algorithm, with the core idea of reducing the overfitting risk of individual decision trees by constructing multiple decision trees. By employing an RF model, significant improvements in prediction accuracy can be achieved, while also mitigating issues such as overfitting, missing values, or multicollinearity. Furthermore, this model can easily address complex qualitative and quantitative problems. A variance inflation factor (VIF) was used for the multicollinearity issue, with a threshold of 5. The multicollinearity issue was not severe in our study ($VIF < 5$). This study utilizes the Python package “sklearn.ensemble.RandomForestRegressor” for modeling. The configuration of the RF model is determined by two adjustable parameters: the number of decision trees (ntree) and the maximum depth of trees (max_depth).

The experimental setup is outlined as follows (Figure 2):

- (1) The environmental factor variables of the sampling points are acquired and combined with heavy metal element contents at the sampling points to form a dataset.
- (2) The dataset is divided into a training set and a validation set in a 7:3 ratio. The training set is used to train RF models with different parameters, and the validation set is utilized for accuracy validation to select the optimal model.
- (3) Construction of the RF model: Utilizing the Bootstrap resampling method, n samples are randomly drawn from the training set to form a new training sample set. With the dataset containing nine feature factors, a random subset of m (where $m \leq 9$) features is selected to form a feature subset. Decision trees are constructed on the new sample set and feature subset, selecting the best splitting attribute during the tree-growing process for node splitting. This process is repeated multiple times to construct decision trees and obtain base estimators. These decision trees are then combined to form the RF.
- (4) Using the “Fishnet” tool in ArcGIS 10.8, a 500 m grid of points was created for the study area. The “Extract Values to Points” tool in ArcGIS 10.8 was then used to extract nine environmental covariates for the grid points.

- (5) The RF model was applied to predict the Cr, Cd, Pb, As, Hg, and Ni contents at each grid point. The experimental results are used to generate spatial distribution prediction maps of heavy metal elements using the “Point to Raster” tool in ArcGIS 10.8.

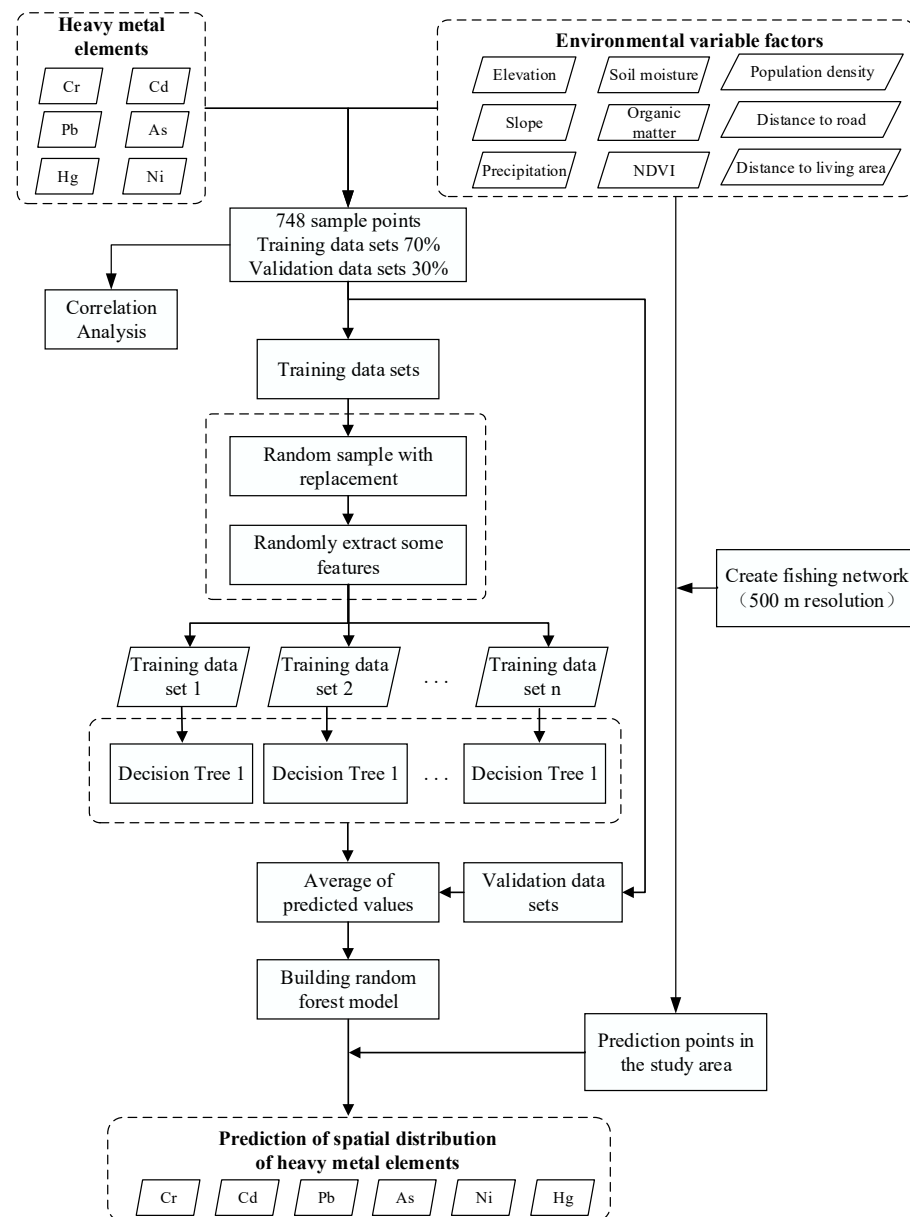


Figure 2. Workflow of prediction of spatial distribution of heavy metal elements based on RF model.

2.4. Parameter Settings for Random Forest Model

Through sequential experiments, the optimal values for the parameters *ntree* and *max_depth* in the RF model can be determined [44]. In Table 2, *max_depth* and *ntree* are selected as 10, 20, 30, and 500, 800, 1000, respectively. Then, three groups of nine experiments were conducted for each setting, and the results are presented in Table 2. To avoid overfitting, this study compares the R^2 values between the training set and validation set, selecting the parameters with the most similar results as the optimal predictive model. The results indicate that when *ntree* is set to 800 and *max_depth* is set to 20, the R^2 values for both the training set and validation set of each metal prediction model are relatively close and relatively large. This suggests that the model exhibits the best stability and robustness with this configuration, providing the most reliable predictions.

Table 2. Selection of optimal parameter values for the RF model.

	ntree	max_Depth	Cr		Cd		Pb	
			Training Set (R ²)	Validation Set (R ²)	Training Set (R ²)	Validation Set (R ²)	Training Set (R ²)	Validation Set (R ²)
Test 1	500	10	0.556	0.507	0.517	0.496	0.521	0.511
	500	20	0.599	0.539	0.518	0.498	0.527	0.521
	500	30	0.599	0.538	0.516	0.497	0.526	0.520
Test 2	800	10	0.559	0.508	0.517	0.499	0.522	0.517
	800	20	0.600	0.542	0.523	0.504	0.535	0.527
	800	30	0.600	0.541	0.522	0.503	0.534	0.526
Test 3	1000	10	0.558	0.507	0.517	0.498	0.521	0.517
	1000	20	0.599	0.540	0.521	0.501	0.534	0.524
	1000	30	0.599	0.540	0.520	0.500	0.533	0.523
	ntree	max_Depth	As		Hg		Ni	
			Training Set (R ²)	Validation Set (R ²)	Training Set (R ²)	Validation Set (R ²)	Training Set (R ²)	Validation Set (R ²)
Test 1	500	10	0.622	0.618	0.605	0.573	0.473	0.454
	500	20	0.627	0.622	0.607	0.579	0.511	0.481
	500	30	0.626	0.621	0.606	0.578	0.510	0.480
Test 2	800	10	0.629	0.625	0.610	0.582	0.472	0.452
	800	20	0.631	0.627	0.612	0.584	0.512	0.482
	800	30	0.630	0.626	0.611	0.583	0.511	0.481
Test 3	1000	10	0.627	0.624	0.609	0.581	0.475	0.453
	1000	20	0.628	0.625	0.610	0.582	0.509	0.481
	1000	30	0.625	0.621	0.610	0.579	0.510	0.480

2.5. Accuracy Evaluation

By employing three parameters, Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and coefficient of determination (R²), the accuracy of the model can be effectively measured. Lower values of MAE and RMSE indicate the higher accuracy of the model, while the R² of the test set is used to evaluate the model's predictive accuracy and generalization ability. The calculation formulas are as follows:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (p_i - o_i)^2} \quad (1)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |p_i - o_i| \quad (2)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (p_i - o_i)^2}{\sum_{i=1}^n (p_i - \hat{o}_i)^2} \quad (3)$$

In these formulas, p_i and o_i represent the predicted and observed values, respectively. $\hat{o}_i$ is the mean of the observed values.

3. Results

3.1. Statistical Characteristics of Heavy Metal Content in Soil

Tables 3 and 4 present the statistical results of the soil heavy metals in the training set and validation set at the sampling points, respectively. Skewness and kurtosis are used to describe the normal distribution of the data, where smaller values typically indicate a closer fit to the normal distribution. The coefficient of variation is a relative statistical measure of

data variability, calculated as the ratio of the standard deviation to the mean. Generally, a smaller coefficient of variation indicates lower data variability, while a larger coefficient of variation suggests higher variability. From the tables, it can be observed that the data on the six heavy metal elements collectively follow a normal distribution. However, the coefficient of variation for Hg is relatively high, indicating its poorer stability.

Table 3. Basic statistical characteristics of soil heavy metals in training set.

Metal	Range (mg/kg)	Mean Value (mg/kg)	Standard Deviation (mg/kg)	Skewness	Kurtosis	Coefficient of Variation (%)	Background Value [45] (mg/kg)
Cr	12.70~144.00	57.01	20.62	1.56	2.96	36.17	57
Cd	0.0031~0.30	0.12	0.059	0.68	0.35	49.17	0.117
Pb	6.87~57.80	23.40	7.22	0.73	1.81	30.85	27.2
As	2.37~21.80	7.65	2.51	1.00	3.57	32.81	6.4
Hg	0.011~1.05	0.072	0.10	4.99	32.06	138.89	0.034
Ni	4.07~82.80	27.74	11.03	1.13	1.34	39.76	24.6

Table 4. Basic statistical characteristics of soil heavy metals in validation set.

Metal	Range (mg/kg)	Mean Value (mg/kg)	Standard Deviation (mg/kg)	Skewness	Kurtosis	Coefficient of Variation (%)	Background Value [45] (mg/kg)
Cr	6.39~147.00	58.35	24.36	1.33	2.07	41.75	57
Cd	0.0161~0.29	0.12	0.060	0.68	0.20	50.00	0.117
Pb	8.25~56.30	23.23	6.82	0.98	2.35	29.36	27.2
As	2.22~21.80	7.24	2.58	1.32	5.50	35.64	6.4
Hg	0.014~0.95	0.058	0.065	4.29	22.68	112.07	0.034
Ni	6.07~81.80	28.03	11.99	0.88	0.24	42.78	24.6

3.2. Correlation Analysis of Factors Affecting Heavy Metals in Soil

Using the “Correlate” module in SPSS Statistics 25 software, a correlation analysis was conducted to investigate the influence of different environmental variables on soil heavy metal content [46]. By conducting a correlation analysis on the data, we can determine the degree of association between different variables, with smaller *p*-values indicating stronger correlations. According to the correlation analysis results presented in Table 5, it is evident that annual average precipitation, elevation, and population density exhibit significant correlations with all heavy metals. Additionally, soil organic matter and the distance from the sampling points to residential areas are correlated with Cd, Pb, As, and Hg. Slope is only correlated with Cr and Pb. Soil humidity and the distance from sampling points to roads show no significant correlation with any of the heavy metals.

The results indicate significant correlations as follows: Cr exhibits significant correlations with annual average precipitation, elevation, slope, and population density. Cd shows significant correlations with soil organic matter, annual average precipitation, elevation, slope, distance from sampling points to residential areas, and population density. Pb demonstrates significant correlations with soil organic matter, annual average precipitation, elevation, slope, distance from sampling points to residential areas, and population density. As displays significant correlations with soil organic matter, annual average precipitation, elevation, distance from sampling points to residential areas, population density, and NDVI. Hg indicates significant correlations with soil organic matter, annual average precipitation, elevation, distance from sampling points to residential areas, distance from sampling points to roads, and population density. Ni shows significant correlations with annual average precipitation, elevation, and population density.

Table 5. Correlation analysis of soil heavy metal contents and environmental variables.

Variable	Cr	Cd	Pb	As	Hg	Ni
Organic Matter	−0.028	0.226 **	0.208 **	0.504 **	0.349 **	−0.030
Precipitation	0.150 **	−0.188 **	−0.218 **	−0.288 **	−0.224 **	0.107 **
Humidity	0.022	−0.055	−0.041	−0.034	−0.062	−0.001
Elevation	0.133 **	−0.233 **	−0.253 **	−0.516 **	−0.374 **	0.081 *
Slope	−0.084 *	0.033	0.099 **	0.044	0.039	−0.056
NDVI	0.041	−0.058	−0.026	−0.056 *	0.026	−0.004
Distance to Living Area	0.012	−0.115 **	−0.102 **	−0.227 **	−0.152 **	0.011
Distance to Road	−0.002	−0.025	−0.007	0.038	0.058 *	0.045
Population Density	−0.159 **	0.144 **	0.229 **	0.233 **	0.299 **	−0.122 **

**, correlation is significant at the 0.01 level (two-tailed). *, correlation is significant at the 0.05 level (two-tailed).

3.3. Importance Analysis of Environmental Variables in Random Forest Model

Utilizing the RF model, the importance of influencing factors in predicting various metals was ranked and analyzed [47]. As depicted in Figure 3, for Cr, the most important factors are population density, annual average precipitation, and elevation. For Cd, the significant factors include annual average precipitation, soil humidity, and soil organic matter. Pb is primarily influenced by annual average precipitation. As is significantly influenced by annual average precipitation and soil organic matter. Hg is notably influenced by annual average precipitation, elevation, and soil organic matter. Ni is significantly affected by population density, annual average precipitation, and slope. This analysis suggests that precipitation may have a diluting effect on soil heavy metals, leading to a significant relationship between precipitation and soil heavy metal content.

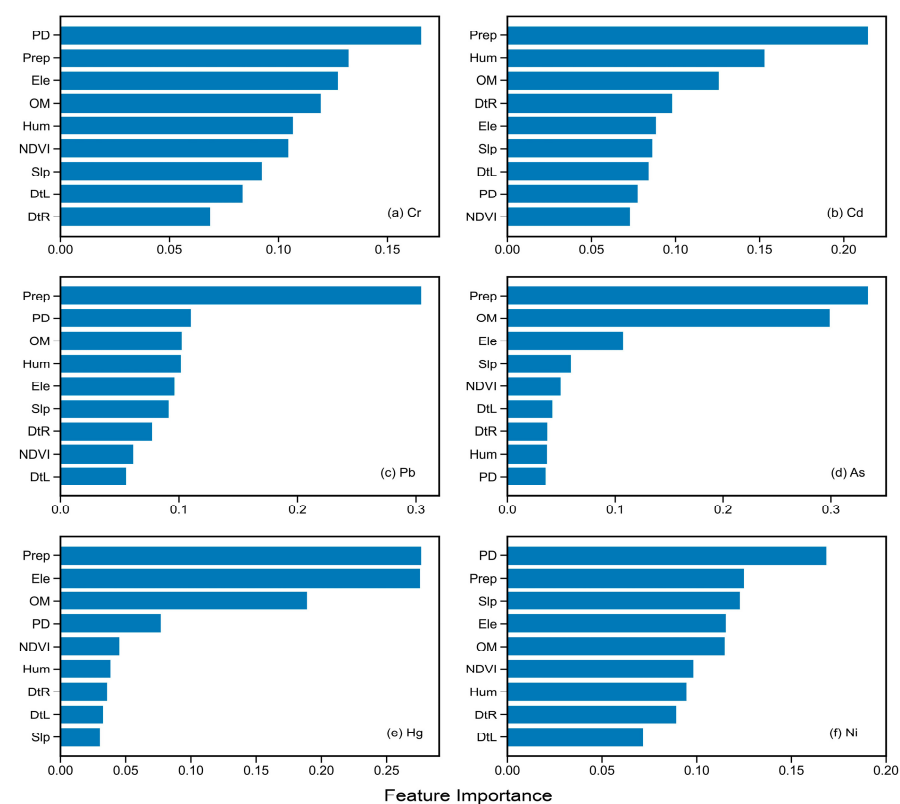


Figure 3. Importance of environmental factors. PD: population density; Prep: precipitation; Ele: elevation; OM: organic matter; Hum: humidity; Slp: slope; DtL: distance to living area; DtR: distance to road.

3.4. Precision Analysis of Heavy Metal Content Prediction Based on Random Forest Model

By calculating evaluation metrics such as the RMSE, MAE, and R^2 , and undergoing multiple rounds of parameter tuning, it was found that the RF model exhibited the best performance. Therefore, we selected this as the most stable model. According to Table 6, (1) the coefficient of determination (R^2) of the RF model closely matches that of the validation set, indicating its ability to effectively avoid overfitting. (2) Based on R^2 values, the prediction performance, from highest to lowest, is as follows: As, Hg, Cr, Pb, Cd, Ni. (3) For soil heavy metal content prediction, the models for As, Hg, Cr, Pb, and Cd exhibit high fitting and generalization capabilities. Both the R^2 values for the training set and the validation set are above 0.5 and are relatively close, indicating stable models. The R^2 value for Ni in the validation set is slightly lower than 0.5, suggesting a relatively poorer prediction accuracy. (4) The analysis of MAE and RMSE reveals that the MAE and RMSE values of the training set are smaller than those of the validation set, indicating the overall higher prediction accuracy of the model. This indicates that, within the study area, the RF model performs well in predicting soil heavy metal content.

Table 6. Precision of the prediction of soil heavy metal.

Element	Training Datasets			Validation Datasets		
	RMSE (mg/kg)	MAE (mg/kg)	R^2	RMSE (mg/kg)	MAE (mg/kg)	R^2
Cr	3.840	10.200	0.600	4.055	11.943	0.542
Cd	0.190	0.028	0.523	0.205	0.032	0.504
Pb	2.047	2.952	0.535	2.164	3.402	0.527
As	1.168	0.945	0.631	1.254	1.023	0.627
Hg	0.241	0.023	0.612	0.205	0.021	0.584
Ni	2.693	5.388	0.512	2.935	6.559	0.482

3.5. Spatial Distribution Prediction of Soil Heavy Metals Based on Random Forest Model

The RF model was utilized to predict the soil heavy metal contents of Cr, Cd, Pb, As, Hg, and Ni in the study area, respectively, and spatial distribution maps were obtained for each soil heavy metal. From Figure 4, it is evident that the distribution of Cd, Pb, As, and Hg in the study area shows higher concentrations in the central and northern regions. Specifically, Cd and As exhibit relatively higher concentrations in the northwest region compared to other areas, while Pb shows higher overall concentrations in the northern region and lower concentrations in the southern region. Hg is predominantly concentrated in a small portion of the central-to-northern region, with lower concentrations in other areas. Conversely, Cr exhibits lower concentrations primarily in the northwest region, with higher concentrations in other areas. Ni shows higher concentrations in the southern region, with distributions in other areas as well. Both Cr and Ni demonstrate higher concentrations in the southern region.

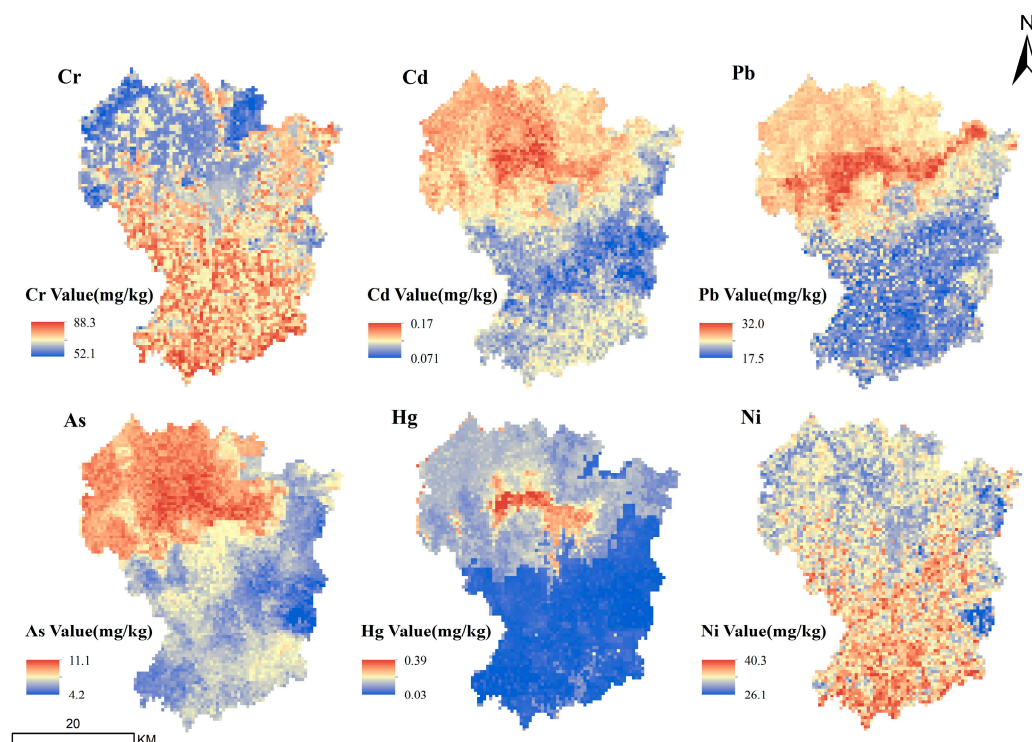


Figure 4. Spatial distribution of soil heavy metals based on RF model.

4. Discussion

Analyzing the sources of high concentrations of different heavy metals is essential for soil pollution prevention. This study revealed significant accumulations of Pb and Hg in the central-to-northern region of the study area, coinciding with the presence of multiple economic development zones and industrial parks [48]. Several studies [49–51] have indicated that industrial emissions are a major source of pollution for Pb and Hg. These pollutants can accumulate in the soil through solid waste, industrial wastewater, or dust deposition from industrial emissions. Therefore, the relatively high concentrations of Pb and Hg in the central to northern parts of the study area are primarily due to industrial production. Furthermore, Cd and As are highly concentrated in the northern region of the study area, which is correlated with agricultural land. Ji [52] showed that there is a significant presence of crops in the northwest part of the study area. Agricultural inputs such as fertilizers and pesticides contain elements such as Cd and As, leading to their accumulation in the soil [53,54]. Our study also found that there are similar distributions of Cr and Ni in the southern parts of the study area, where human activities are relatively limited. Research by Wang [55] has demonstrated a strong homogeneity between Cr and Ni, indicating that their primary sources are natural sources.

Although feature selection is regarded as an important step in the RF model [56], all environmental factors selected in our study were used to train the model. The current study has shown that all these factors are correlated with soil heavy metals. We think the existing factors can provide more comprehensive environmental information, which is important for the spatial prediction of soil heavy metal. Precipitation and population density highly relate to soil heavy metal contents, a fact which is supported by the study of Zhu [32]. Agricultural and industrial activities in human settlements are important factors affecting soil heavy metal content, thus leading to a significant correlation between population density and soil heavy metal content. Furthermore, there are additional environmental factors that may impact soil heavy metal content. For example, slope direction, various bands of remote sensing imagery, and the distance from sampling points to rivers were included as environmental variables for the prediction of As's distribution in soil [57]. Liu [58] took

into account the pH when assessing soil heavy metal pollution, while Jiang [59] considered PM2.5, GDP, and the duration of nighttime lighting in the Yangtze River region.

Geographically Weighted Regression (GWR), Support Vector Machine (SVM), and other machine learning algorithms can also be used to predict soil heavy metal contents [24,60], but the RF model was chosen in our study due to its strong robustness, resistance to overfitting, and insensitivity to multicollinearity [20]. Ma [24] also used an RF model to predict soil heavy metal hyperspectral data, showing that the RF algorithm can achieve higher accuracy results compared to the ELM (Extreme Learning Machine) and SVM algorithms.

Although the spatial distributions of various soil heavy metals were obtained in our study, there are still some limitations. The R^2 in our study is common in the spatial prediction of soil heavy metals using the RF model [24,61,62], especially in large-scale studies. But our prediction accuracy can be further improved in some respects. First, more environmental factors can be included to provide additional information and enhance prediction accuracy. Azizi [63] proved that increasing the number of environmental factors continuously can raise the R^2 from 0.24 to 0.63 in soil heavy metal prediction. Second, more parameter optimization experiments can enhance prediction accuracy. Additionally, although excessive levels of heavy metals such as Cuprum (Cu) and Zinc (Zn) can also impact both soil ecosystems and human health, these elements are essential for plant growth and maintaining human physiological functions [64]. Therefore, our study did not conduct spatial distribution predictions for these soil heavy metal elements. In subsequent experiments, more environmental factors will be considered, and various prediction methods will be compared for mapping the spatial distribution of soil heavy metals.

5. Conclusions

The use of the RF model in this study for predicting the spatial distribution of soil heavy metals holds significant importance in improving prediction accuracy. The research findings are as follows:

- (1) The modeling results of the RF model indicate that environmental variables play a significant role in explaining the variations in the soil heavy metal content of Cr, Cd, As, Pb, Hg, and Ni within the study area. The close similarity between the R^2 values of the training and validation sets suggests that the RF model exhibits reduced overfitting issues and higher stability in predicting the spatial distribution of heavy metals in the soil within the study area. Therefore, the RF model demonstrates a favorable performance in predicting the spatial distribution of soil heavy metal content.
- (2) For the soil heavy metal content in the study area, annual precipitation and population density are identified as major influencing factors. Specifically, Cd and Hg are primarily water-soluble components, while Cr, As, Pb, and Ni mainly exist in insoluble forms in water. Therefore, precipitation plays a significant role in soil heavy metal dynamics due to its dilution and dissolution effects. Consequently, there is a substantial relationship between precipitation and soil heavy metal content. Additionally, human activities are significant factors influencing soil heavy metal content. Hence, there is a notable correlation between population density and soil heavy metal content.
- (3) Based on the predicted distribution maps and the natural and social environmental conditions of the study area, it is evident that industrial activities can lead to elevated levels of Pb and Hg in the soil. The use of agricultural products in agricultural production can result in increased levels of Cd and As in the soil. Cr and Ni are primarily influenced by natural environmental factors and precipitation, with less influence from human activities.

From the research findings, it is evident that the RF model can accurately predict the distribution of soil heavy metals. Subsequently, it would be beneficial to explore the integration of the RF model with other machine learning methods for predicting the spatial distribution of soil heavy metals. This integration could potentially yield even more precise results, thereby providing more reliable insights for environmental management and protection.

Author Contributions: Methodology, S.N.; Formal analysis, S.N.; Resources, H.C.; Data curation, X.S.; Writing—original draft, S.N.; Supervision, Y.A.; Funding acquisition, H.C. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the National Natural Science Foundation of China (42101430).

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

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

References

- Gao, Q.; Yu, M.; Liu, Y.H.; Xu, H.M.; Xu, X. Modeling interplay between regional net ecosystem carbon balance and soil erosion for a crop-pasture region. *J. Geophys. Res.-Biogeosci.* **2007**, *112*. [\[CrossRef\]](#)
- Cuevas, J.; Daliakopoulos, I.N.; Del Moral, F.; Hueso, J.J.; Tsanis, I.K. A Review of Soil-Improving Cropping Systems for Soil Salinization. *Agronomy* **2019**, *9*, 295. [\[CrossRef\]](#)
- Li, G.; Sun, G.X.; Ren, Y.; Luo, X.S.; Zhu, Y.G. Urban soil and human health: A review. *Eur. J. Soil Sci.* **2018**, *69*, 196–215. [\[CrossRef\]](#)
- Chen, W.P.; Yang, Y.; Xie, T.; Wang, M.E.; Peng, C.; Wang, R.D. Challenges and Countermeasures for Heavy Metal Pollution Control in Farmlands of China. *Acta Pedol. Sin.* **2018**, *55*, 261–272.
- Jiao, S.; Chen, Z.H.; Yu, A.H.; Chen, H.H. Evaluation of the heavy metal pollution ecological risk in topsoil: A case study from Nanjing, China. *Environ. Earth Sci.* **2022**, *81*, 532. [\[CrossRef\]](#)
- Alyemeni, M.N.; Almohisen, I.A.A. Traffic and industrial activities around Riyadh cause the accumulation of heavy metals in legumes: A case study. *Saudi J. Biol. Sci.* **2014**, *21*, 167–172. [\[CrossRef\]](#) [\[PubMed\]](#)
- Cao, M.Z.; Zhu, W.J.; Hong, L.D.; Wang, W.P.; Yao, Y.L.; Zhu, F.X.; Hong, C.L.; He, S.Y. Assessing Pb-Cr Pollution Thresholds for Ecological Risk and Potential Health Risk in Selected Several Kinds of Rice. *Toxics* **2022**, *10*, 645. [\[CrossRef\]](#)
- Chen, H.H.; Wang, L.; Yu, A.H. Evaluation of heavy metal pollution in the soil surface of adaptive FCM—Taking the Nanjing refinery and its living area as an example. *China Environ. Sci.* **2022**, *42*, 5239–5245.
- Yaashikaa, P.R.; Kumar, P.S.; Jeevanantham, S.; Saravanan, R. A review on bioremediation approach for heavy metal detoxification and accumulation in plants. *Environ. Pollut.* **2022**, *301*, 119035. [\[CrossRef\]](#)
- Burges, A.; Epelde, L.; Garbisu, C. Impact of repeated single-metal and multi-metal pollution events on soil quality. *Chemosphere* **2015**, *120*, 8–15. [\[CrossRef\]](#)
- Li, W.J.; Yin, Z.X.; Yue, B.; Gao, T.P.; Chang, G.H. Distribution and Risk Assessment of Some Heavy Metal Elements in the Contaminated soil from Baiyin City, Gansu Province. *IOP Conf. Ser. Earth Environ. Sci.* **2020**, *568*, 012044. [\[CrossRef\]](#)
- Cheng, H.; Shen, R.L.; Chen, Y.Y.; Wan, Q.J.; Shi, T.Z.; Wang, J.J.; Wan, Y.; Hong, Y.S.; Li, X.C. Estimating heavy metal concentrations in suburban soils with reflectance spectroscopy. *Geoderma* **2019**, *336*, 59–67. [\[CrossRef\]](#)
- Xu, X.X.; Shi, F.D.; Zhu, J.J. Analyzing the critical factors influencing residents' willingness to pay for old residential neighborhoods renewal: Insights from Nanjing, China. *Environ. Dev. Sustain.* **2024**. [\[CrossRef\]](#)
- Yang, Q.; Zheng, J.Z.; Zhu, H.C. Influence of spatiotemporal change of temperature and rainfall on major grain yields in southern Jiangsu Province, China. *Glob. Ecol. Conserv.* **2020**, *21*, e00818. [\[CrossRef\]](#)
- Han, X.; Wu, H.; Li, Q.; Cai, W.; Hu, S. Assessment of heavy metal accumulation and potential risks in surface sediment of estuary area: A case study of Dagou river. *Mar. Environ. Res.* **2024**, *196*, 106416. [\[CrossRef\]](#) [\[PubMed\]](#)
- Jin, Z.; Lv, J.S. Comparison of the accuracy of spatial prediction for heavy metals in regional soils based on machine learning models. *Geogr. Res.* **2022**, *41*, 1731–1747.
- Mansuy, N.; Thiffault, E.; Paré, D.; Bernier, P.; Guindon, L.; Villemaire, P.; Poirier, V.; Beaudoin, A. Digital mapping of soil properties in Canadian managed forests at 250 m of resolution using the *k*-nearest neighbor method. *Geoderma* **2014**, *235*, 59–73. [\[CrossRef\]](#)
- Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **2008**, *28*, 1–26. [\[CrossRef\]](#)
- Henderson, B.L.; Bui, E.N.; Moran, C.J.; Simon, D.A.P. Australia-wide predictions of soil properties using, decision trees. *Geoderma* **2005**, *124*, 383–398. [\[CrossRef\]](#)
- Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
- Dharumarajan, S.; Hegde, R.; Singh, S.K. Spatial prediction of major soil properties using Random Forest techniques—A case study in semi-arid tropics of South India. *Geoderma Reg.* **2017**, *10*, 154–162. [\[CrossRef\]](#)
- Chagas CD, S.; Junior WD, C.; Bhering, S.B.; Filho, B.C. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. *Catena* **2016**, *139*, 232–240. [\[CrossRef\]](#)
- Guo, P.T.; Li, M.F.; Luo, W.; Lin, Q.H.; Tang, Q.F.; Liu, Z.W. Prediction of soil total nitrogen for rubber plantation at regional scale based on environmental variables and random forest approach. *Trans. Chin. Soc. Agric. Eng.* **2015**, *31*, 194–202.
- Ma, W.; Tan, K.; Du, P. Predicting soil heavy metal based on Random Forest model. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 4331–4334.
- Tan, K.; Wang, H.; Chen, L.; Du, Q.; Du, P.J.; Pan, C.C. Estimation of the spatial distribution of heavy metal in agricultural soils using airborne hyperspectral imaging and random forest. *J. Hazard. Mater.* **2020**, *382*, 120987. [\[CrossRef\]](#) [\[PubMed\]](#)

26. Xu, J.; Xiao, P. Influence factor analysis of soil heavy metal based on categorical regression. *Int. J. Environ. Sci. Technol.* **2022**, *19*, 7373–7386. [[CrossRef](#)]
27. Yang, J.; Wang, J.Y.; Qiao, P.W.; Zheng, Y.M.; Yang, J.X.; Chen, T.B.; Lei, M.; Wan, X.M.; Zhou, X.Y. Identifying factors that influence soil heavy metals by using categorical regression analysis: A case study in Beijing, China. *Front. Environ. Sci. Eng.* **2020**, *14*, 37. [[CrossRef](#)]
28. Shin, H.; Yu, J.; Wang, L.; Jeong, Y.; Kim, J. Spectral Interference of Heavy Metal Contamination on Spectral Signals of Moisture Content for Heavy Metal Contaminated Soils. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 2266–2275. [[CrossRef](#)]
29. Kou, B.; Yuan, Y.; Zhu, X.; Ke, Y.; Wang, H.; Yu, T.Q.; Tan, W.B. Effect of soil organic matter-mediated electron transfer on heavy metal remediation: Current status and perspectives. *Sci. Total Environ.* **2024**, *917*, 170451. [[CrossRef](#)]
30. Zeng, F.; Ali, S.; Zhang, H.; Ouyang, Y.N.; Qiu, B.Y.; Wu, F.B.; Zhang, G.P. The influence of pH and organic matter content in paddy soil on heavy metal availability and their uptake by rice plants. *Environ. Pollut.* **2011**, *159*, 84–91. [[CrossRef](#)]
31. Cao, J.; Xie, C.Y.; Hou, Z.R. Transport patterns and numerical simulation of heavy metal pollutants in soils of lead–zinc ore mines. *J. Mt. Sci.* **2021**, *18*, 2345–2356. [[CrossRef](#)]
32. Zhu, P.; Cui, S.S.; Li, Z.T.; Zhu, X.T.; He, J.L.; Tan, H. Influence of Atmospheric Precipitation on the Release of Cadmium from High Background Soils in Karst Areas of Guizhou. *Ecol. Environ.* **2021**, *30*, 2213–2222.
33. Ling, X.D.; Wang, L.Q.; Zhao, K.L.; Fu, W.J.; Ye, Z.Q.; Ding, L.Z. Spatial distribution characteristics of soil available nutrients in hickory plantation based on random forest method. *Acta Ecol. Sin.* **2024**, *44*, 662–675.
34. He, M.Y.; Dong, J.B.; Jin, Z.; Liu, C.Y.; Xiao, J.; Zhang, F.; Sun, H.; Zhao, Z.Q.; Gou, L.F.; Liu, W.G.; et al. Pedogenic processes in loess-paleosol sediments: Clues from Li isotopes of leachate in Luochuan loess. *Geochim. Et Cosmochim. Acta* **2021**, *299*, 151–162. [[CrossRef](#)]
35. Bai, B.; Xu, T.; Nie, Q.; Li, P.P. Temperature-driven migration of heavy metal Pb²⁺ along with moisture movement in unsaturated soils. *Int. J. Heat Mass Transf.* **2020**, *153*, 119573. [[CrossRef](#)]
36. Dai, Q.Q.; Xu, M.J.; Zhuang, S.Y.; Chen, D.F. Study on Factors Influencing Heavy Metal of Farmland Soils Based on Geographical Detector in Fengqiu County. *Soils* **2022**, *54*, 564–571.
37. Guo, G.H.; Lei, M.; Chen, T.B.; Song, B.; Li, X.Y. Effect of road traffic on heavy metals in road dusts and roadside soils. *Acta Sci. Circumstantiae* **2008**, *28*, 1937–1945.
38. Liu, P.J.; Wu, K.N.; Luo, N. Potential Risk Factors Identification of Heavy Metals Spatial Variation in Typical Agricultural Land Topsoil of Taihu Basin. *Resour. Environ. Yangtze Basin* **2020**, *29*, 609–622.
39. GB 15618-2018; Soil environmental quality Risk control standard for soil contamination of agricultural land. Standardization Administration: Beijing, China, 2018.
40. Ding, Y.X.; Peng, S.Z. Spatiotemporal Trends and Attribution of Drought across China from 1901–2100. *Sustainability* **2020**, *12*, 477. [[CrossRef](#)]
41. Li, Q.L.; Shi, G.S.; Shangguan, W.; Nourani, V.; Li, J.D.; Li, L.; Huang, F.N.; Zhang, Y.; Wang, C.Y.; Wang, D.G.; et al. A 1 km daily soil moisture dataset over China using in situ measurement and machine learning. *Earth Syst. Sci. Data* **2022**, *14*, 5267–5286. [[CrossRef](#)]
42. Yang, J.; Dong, J.; Xiao, X.; Dai, J.; Wu, C.; Xia, J.; Zhao, G.; Zhao, M.; Li, Z.; Zhang, Y.; et al. Divergent shifts in peak photosynthesis timing of temperate and alpine grasslands in China. *Remote Sens. Environ.* **2019**, *233*, 111395. [[CrossRef](#)]
43. Tatem, A.J. WorldPop, open data for spatial demography. *Sci. Data* **2017**, *4*, 170004. [[CrossRef](#)] [[PubMed](#)]
44. Lu, H.L.; Zhao, M.S.; Liu, B.Y.; Zhang, P.; Lu, L.M. Spatial Prediction of Soil Properties Based on Random Forest Model in Anhui Province. *Soils* **2019**, *51*, 602–608.
45. Zhao, C.; Wang, Q.; Dai, J.P. Investigation and analysis of background value of heavy metals in soil of Shandong Province. *Environ. Prot. Sci.* **2021**, *47*, 117–121.
46. Song, J.Q.; Zhu, Q.; Jiang, X.S.; Zhao, H.Y.; Liang, Y.H.; Luo, Y.X.; Wang, Q.; Zhao, L.L. GIS-Based Heavy Metals Risk Assessment of Agricultural Soils—A Case Study of Baguazhou, Nanjing. *Acta Pedol. Sin.* **2017**, *54*, 81–91.
47. Zhou, Y.F.; Xie, B.L.; Li, M.S. Mapping regional forest aboveground biomass from random forest Co-Kriging approach: A case study from north Guangdong. *J. Nanjing For. Univ. (Nat. Sci. Ed.)* **2023**, *48*, 169–178.
48. Chen, Y.C.; Duan, W.T.; Chi, Y.H.; Li, P.; Liu, H. Study on County Village Type Identification Under the Background of Urban–rural Integration Development: A Case Study of Zhaoyuan City in Shandong Province. *Urban Dev. Stud.* **2022**, *29*, 28–37.
49. Li, Z.Y.; Ma, Z.W.; Van Der Kuijp, T.J.; Yuan, Z.W.; Huang, L. A review of soil heavy metal pollution from mines in China: Pollution and health risk assessment. *Sci. Total Environ.* **2014**, *468*, 843–853. [[CrossRef](#)] [[PubMed](#)]
50. Sun, X.F.; Zhang, L.X.; Dong, Y.L.; Zhu, L.Y.; Wang, Z.; Lu, J.T. Source Apportionment and Spatial Distribution Simulation of Heavy Metals in a Typical Petrochemical Industrial City. *Environ. Sci.* **2021**, *42*, 1093–1104.
51. Li, Y.L.; Chen, W.P.; Yang, Y.; Wang, T.Q.; Liu, C.F.; Cai, B. Heavy metal pollution characteristics and comprehensive risk evaluation of farmland across the eastern plain of Jiyuan city. *Acta Sci. Circumstantiae* **2020**, *40*, 2229–2236.
52. Ji, F.H.; Liu, J.; Wang, L.M. Summary of Remote Sensing Algorithm in Crop Type Identification and Its Application Based on Gaofen Satellites. *Chin. J. Agric. Resour. Reg. Plan.* **2021**, *42*, 254–268.
53. Cheng, J.; Yuan, X.Y.; Zhang, H.Y.; Mao, Z.Q.; Zhu, H.; Wang, Y.M.; Li, J.Z. Characteristics of Heavy Metal Pollution in Soils of Yunnan-Guizhou Phosphate Ore Areas and Their Effects on Quality of Agricultural Products. *J. Ecol. Rural Environ.* **2021**, *37*, 636–643.

54. Huang, H.B.; Lin, C.Q.; Hu, G.R.; Yu, R.L.; Hao, C.L.; Chen, F.H. Source Appointment of Heavy Metals in Agricultural Soils of the Jiulong River Basin Based on Positive Matrix Factorization. *Environ. Sci.* **2020**, *41*, 430–437.
55. Wang, M.E.; Peng, C.; Chen, W.P. Impacts of Industrial Zone in Arid Area in Ningxia Province on the Accumulation of Heavy Metals in Agricultural Soils. *Environ. Sci.* **2016**, *37*, 3532–3539.
56. Belgiu, M.; Drăguț, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31. [[CrossRef](#)]
57. Shi, G.; Liu, G.; Zhao, L.; Su, Y.Q.; Bi, R.T. Prediction of arsenic for farmland soil based on multi source environmental data and random forest model. *Acta Sci. Circumstantiae* **2020**, *40*, 2993–3000.
58. Liu, Q.; Chen, W.; Wang, B.; Wang, S.; Liu, Z.Z.; Zhang, N.M.; Li, B. Contaminant Characteristics and Health Risk Assessment of Heavy Metals in Soils from Lead-Zincs Melting Plant in Huize County, Yunnan Province, China. *J. Agric. Resour. Environ.* **2024**, *33*, 221.
59. Jiang, Y.F.; Huang, M.X.; Chen, X.Y.; Wang, Z.G.; Xiao, L.J.; Xu, K.; Zhang, S.; Wang, M.M.; Xu, Z.; Shi, Z. Identification and risk prediction of potentially contaminated sites in the Yangtze River Delta. *Sci. Total Environ.* **2022**, *815*, 151982. [[CrossRef](#)] [[PubMed](#)]
60. Li, T.Y.; Jia, W.W.; Sun, Y.M.; Wang, H.Z.; Ma, S.Y. Analysis of Spatial Distribution of Korean Pines in Liangshui Nature Reserve Based on the Geographically Weighted Regression Model. *For. Eng.* **2024**, *40*, 47–59.
61. Wang, Q.; Xie, Z.; Li, F. Using ensemble models to identify and apportion heavy metal pollution sources in agricultural soils on a local scale. *Environ. Pollut.* **2015**, *206*, 227–235. [[CrossRef](#)]
62. 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. [[CrossRef](#)]
63. Azizi, K.; Ayoubi, S.; Nabiollahi, K.; Garosi, Y.; Gislum, R. Predicting heavy metal contents by applying machine learning approaches and environmental covariates in west of Iran. *J. Geochem. Explor.* **2022**, *233*, 106921. [[CrossRef](#)]
64. Liu, W.R.; Zeng, D.; She, L.; Su, W.X.; He, D.C.; Wu, G.Y.; Ma, X.R.; Jiang, S.; Jiang, C.H.; Ying, G.G. Comparisons of pollution characteristics, emission situations, and mass loads for heavy metals in the manures of different livestock and poultry in China. *Sci. Total Environ.* **2020**, *734*, 139023. [[CrossRef](#)] [[PubMed](#)]

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.