sustainability

MDPI

Article

Retrieval of Soil Heavy Metal Content for Environment Monitoring in Mining Area via Transfer Learning

Yun Yang [1,2], Qinfang Cui [3], Rongjie Cheng [1,*], Aidi Huo [4,*] and Yanting Wang [1]

[1] College of Geological Engineering and Geomatics, Chang'an University, No.126, Yanta Road, Xi'an 710054, China; yangyunbox@chd.edu.cn (Y.Y.)
[2] Key Laboratory of Formation Mechanism and Prevention and Control of Mine Geological Disasters, Ministry of Natural Resources, No.126, Yanta Road, Xi'an 710054, China
[3] Piesat International Information Technology Co., Ltd., No.532, Shenzhou Third Road, Xi'an 710199, China
[4] School of Water and Environment, Chang'an University, No.126, Yanta Road, Xi'an 710054, China
* Correspondence: sher_221@163.com (R.C.); huoaidi@chd.edu.cn (A.H.)

Abstract: Monitoring environmental pollution sources is an ongoing issue that must be addressed to reduce risks to public health, food safety, and the environment. However, retrieving topsoil heavy metal content at a low cost for environmental monitoring in mining areas is challenging. Therefore, this study proposes a network model based on transfer learning theory and a back propagation (BP) network optimized by a genetic algorithm (GA), taking the Daxigou mining area in Shaanxi Province, China, as a case study. Firstly, visible and near-infrared spectrum data from Landsat8 satellite images, digital elevation models, and geochemical data from field-collected soil samples were used to extract environmental factor candidates indicating the content and spatial distribution of certain heavy metals, including copper (Cu) and lead (Pb). Secondly, each element was correlated with environmental factors and a multicollinearity test was performed to determine the optimal factor set. Then, the BP network optimized by GA was pre-trained with sample data collected in 2017 and retrained with minimal sample data from 2019 using the parameter transfer learning method, allowing spatial distribution mapping of the Cu and Pb content in topsoil of the Daxigou mining area in 2019. From the validation results using field-collected data, the root mean square error (RMSE) and mean relative error (MRE) values using the proposed model, respectively, reduced by 4.688 mg/kg and 1.533 mg/kg for Cu and reduced by 1.586 mg/kg and 1.232 mg/kg for Pb compared to the traditional GA-BP model. Thus, conclusions can be drawn that our proposed Tr-GA-BP network performs well, requiring 16 training samples collected in 2019. In addition, the content of Cu is the highest; Pb is the second highest in the study area. Both of them were spatially distributed mainly in the exploitation, slag stacking, roadside, etc., consistent with field investigation results.

Keywords: soil heavy metal; multispectral remote sensing; neural network; transfer learning; mining area

Citation: Yang, Y.; Cui, Q.; Cheng, R.; Huo, A.; Wang, Y. Retrieval of Soil Heavy Metal Content for Environment Monitoring in Mining Area via Transfer Learning. *Sustainability* **2023**, *15*, 11765. https://doi.org/10.3390/su151511765

Academic Editors: Lidija Ćurković and Mihone Kerolli Mustafa

Received: 24 June 2023
Revised: 21 July 2023
Accepted: 24 July 2023
Published: 31 July 2023

1. Introduction

Insufficiently treated wastewater, dust, and municipal and industrial waste, especially from mining activities, have caused an increase in the content of heavy metals in soil and groundwater over the past decades. There has been a progressive degradation of the environment and a serious threat to food safety and public health [1,2]. One of the main factors of negative human impact on the natural environment is the release of heavy metals, which pose a serious threat to living organisms [3]. This is an unfavorable and dangerous phenomenon because compounds of such elements as copper, chromium, cadmium, or lead are not biodegradable and accumulate in living organisms, thus passing into the trophic chain and posing a threat to human health and even life. Therefore, efficiently investigating and monitoring the circumstances of soil threatened by heavy metal, especially in mining

areas, for pollution control, ecological protection, and public health, is a key scientific problem currently faced by China [4].

To address these issues, scholars have performed many studies. For example, scholars [5,6] used visible/near-infrared data measured by a hand-held spectrometer to construct a regression model for the retrieval of the soil heavy metals in a farmland area; Gu et al. [7] used laser-induced break-down spectroscopy (LIBS) measurement technology with a combination of laboratory analysis data from soil samples to map the spatial distribution of the soil heavy metal content. However, such site-by-site measurement technology has a high cost for measurement work in the field for large-scale pollution investigations.

Hyperspectral remote sensing imagery has been proven to be effective for directly or indirectly reflecting the characteristics of soil- and vegetation-covered surfaces at a large scale and short period. Tan et al. [8] proposed estimating the spatial distribution of heavy metal in agricultural soils using airborne hyperspectral imaging and random forest. Zhang et al. [9] made contributions on the issue in soils of potentially polluted sites based on unmanned aerial vehicle (UAV) hyperspectral imagery. The authors [10,11] summarized previous research on soil heavy metal content estimation using different data sources and analyzed the ongoing challenges and existing issues.

However, the cost of acquiring hyperspectral imagery with medium or high spatial resolution is usually high. Therefore, some scholars, such as Peng et al. [12], utilized Landsat8 multi-spectral imagery, spectral indices, and auxiliary environmental variables to model and map the spatial distribution of heavy metals in Qatari soils. Investigating these studies, the selection of factors that could be effective for the retrieval of soil heavy metal content is a focus of research so as to make up the insufficiency of spectral information of multispectral imagery.

As stated by Wang et al. [10], soil is a complex mixed system composed of many components and affected by a large number of environmental variables; thus, it is difficult to explicitly determine the relationship between observations and the content of soil heavy metal from physical and chemical mechanisms alone. Therefore, choosing an effective regression model also is vital to improve the precision of the retrieval of soil heavy metal content. According to the previous related publications, e.g., [9,13–15], statistical and machine learning models such as partial least squares regression (PLSR), support vector regression (SVR), M5 model tree, extreme learning machines, random forest, or back propagation are popular for modeling the complex quantitative retrieval problems due to their advantages of simple structure and low training cost compared with popular deep learning networks.

However, it is not enough to depend on common factors and popular estimation models; more factors reflecting the relationship of the adsorption or occurrence among organic matter, clay minerals, and other soil parameters should be incorporated. Authors, e.g., [16–19], made contributions on the issue. The studies [6,8,17] analyzed the reflectance and adsorption mechanism of soil heavy metals based on spectral characteristics of multispectral or hyperspectral data. From the standpoint of the interaction or occurrence relationships, scholars [16,18,19] focused on the interaction between heavy metals and soil constituents such as organic matter, moisture content, and metallic oxides as important factors influencing the potential for soil, crop, and ground water pollution by heavy metals. Considering the interaction between heavy metal and soil constituents to indirectly deduce the content of the soil heavy metals is becoming a novel path to solve the problem.

The previous studies have made substantial contributions to the development of research on the retrieval of heavy metal content in topsoil from different viewpoints. Nevertheless, there is an issue that should be considered, that is, the generalization ability of statistical regression and machine learning models is weak for other similar scenes or the same scene at different times if there are limited training samples. So, the difficulties of reducing the data collection cost and updating the operating mode for soil heavy metal content retrieval should be solved.

The transfer learning theory framework was introduced to solve the few-shot learning problem, which does not require large-scale training data and has low learning costs. It can reduce the data-collection cost by transferring training data from source to target. The theory has a very broad application in fields with limited data volume. In the field of soil contamination monitoring, it has high cost to collect a great number of training samples in complex terrain and land cover; also, training data for parameters of machine learning models are hardly available to the public, unlike in the field of target detection and classification. Therefore, it is of great advantage to introduce transfer learning theory framework in the application of soil heavy metal retrieval, which can reduce the data-collection cost by transferring sample data or prior information from one task to another task.

However, only limited studies, such as [20], have been investigated on the transfer ability of quantitative retrieval models from one scene to the others for soil heavy metal (Pb and Zn) pollution mapping. In previous work focusing on the quantitative inversion problems in soil contamination monitoring using the transfer learning framework, some of the existing problems are what can be transferred from source domain to target domain and how we can transfer data or knowledge from source domain to target domain under the conditions of insufficient training samples. One of the great difficulties is how we can avoid negative transfer of training samples from source domain to target domain while keeping high regression precision.

In this context, our study proposes an innovative quantitative retrieval method by combining a GA-BP neural network with a parameter transfer learning strategy in order to map the spatial distribution of soil heavy metal in the same area but in different periods, considering Daxigou siderite in Shaanxi in China as a case study. The purpose of this study is to evaluate soil pollution and harm to human health from heavy metal in the mining area and surroundings and to provide assistance for decision making for land degradation control, ecological environment protection, and restoration.

2. Materials

2.1. The Study Area

The Daxigou mining area is in Xiaoling Town, Zhashui County, Shangluo City, Shaanxi Province, China, with a designated area of 4.33 km^2, as shown in Figure 1. It is the largest siderite operation in Shaanxi, accounting for 47.6% of the total iron ore reserves. In 1982, the Northwest Metallurgical Geological Exploration Company found that the Daxigou–Yindongzi deposit is rich in copper, lead, zinc, silver, etc., in and around the mining area [21]. Mining in the area officially began in 1988, and open-pit mining has mainly been used since 2007.

In the study, the mining area and its surroundings covering 39 km^2 areas, where the content of copper and lead is relatively high, were considered as our study case. There are mainly medium gullies and low gullies, with large elevation difference and a complex topography in the area, which belongs to the structural erosion landform. The main land use categories are mining area, cultivated land, forestland, grassland, industrial and mining facilities, and residential area. Mining activities have caused heavy metal pollution and ecological and geological environment damage since 1988 [22]. Therefore, it is of great necessity to investigate and regularly monitor the heavy metal pollution in the area.

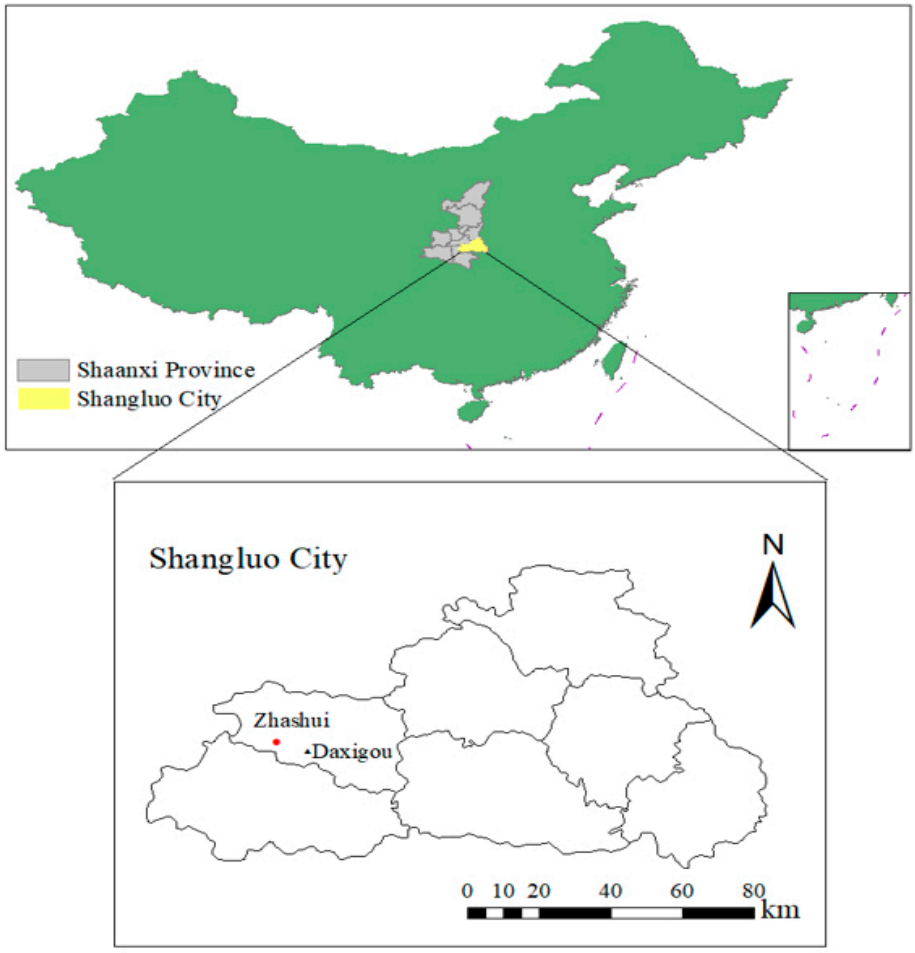


Figure 1. Geographical position of the study area in China.

2.2. Data Preparation

2.2.1. Soil Sample Collection

According to the topography, geomorphic characteristics, and land use types in the study area, soil samples were obtained along the three main ridge lines considering their representativeness and the uniform distribution of the sample points. The sampling points were mainly distributed both in the middle of the hillsides where it was possible to reach and in sites close to the valley. The site distribution with a plum blossom shape was designed within a range of 30 m × 30 m from the sampling point center. The accumulation of heavy metals at the bottom of the slope usually was high due to scouring, where the sampling depth is approximately 20–30 cm, while the sampling depth on the middle of the slope was approximately 10–20 cm.

The soil within the 30 m × 30 m coverage was mixed equally and then 1 kg of soil was removed and placed into the sample package. Simultaneously, WGS84 coordinates of the central point of the sampling area were recorded for each sample. In addition, the soil attributes and its environment observations, including the pressure, position, and the land use category in each site, were recorded.

According to the scheme above, 44 and 43 total soil samples were collected from the field by professional technicians in October 2017 and October 2019, respectively. The sampling site distribution and soil samples to be analyzed in the laboratory are shown in Figure 2a,b.

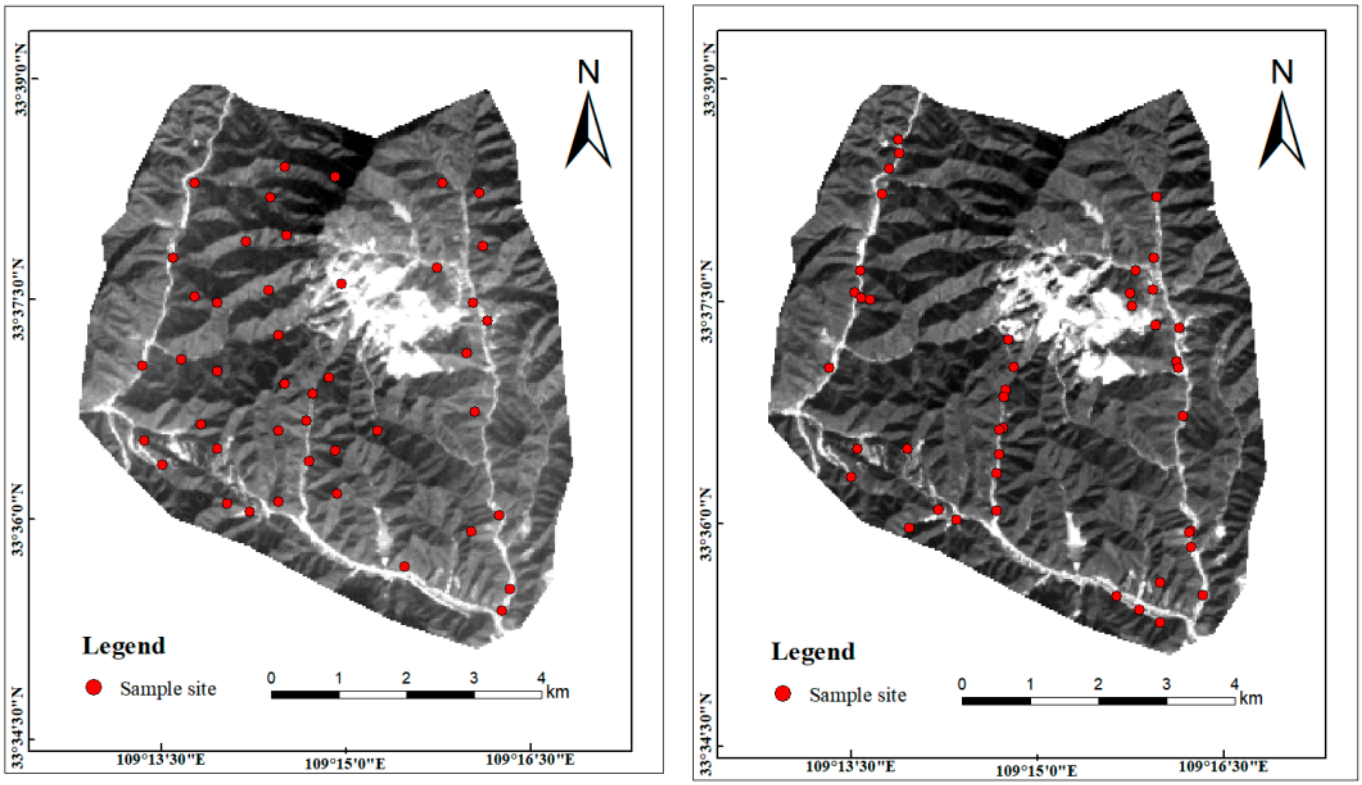


Figure 2. The distribution of sampling sites in the study area. (**a**) The distribution of sample sites over Landsat8 image in 2017; (**b**) the distribution of sample sites over Landsat8 image in 2019.

2.2.2. Soil Sample Analysis and Preprocessing

Each sample collected in 2017 and in 2019 was crushed to remove the remains of animals and plants and then dried, followed by a screening operation with a nylon screen for laboratory analysis.

To determine the interesting heavy metals in the study area, a mixed sample was formed from 44 samples collected in 2017 and the content of eight popular heavy metals (Hg, As, Cd, Cu, Ni, Zn, Pb, and Cr) in soil pollution were analyzed using professional instrument and detection methods, determined by China National Environmental Monitoring Center, in a laboratory of environmental testing center of Guolian Quality Inspection Technology Co., Ltd. in Xi'an, China. Among them, Cu and Pb content were measured using flame atomic absorption spectrophotometry.

By comparing the detected values of each element with their reference values published by State Environmental Protection Administration of China [23], according to the degree to which the detected values exceeded the reference and combined with the enrichment degree of heavy metals in Daxigou–Yindongzi polymetallic ore deposit [21] and the total cost of soil samples to be analyzed in the laboratory, Cu and Pb were determined to be the elements of interest in this study.

The histogram analysis of Cu and Pb in all of the soil samples in 2017 and in 2019 was performed, respectively. From the histograms of the content of the two elements of interest, the content of the Cu and Pb had abnormal values, which would affect the accuracy of the estimation model; therefore, the maximum abnormal values were eliminated. Consequently, the effective number of samples of Cu and Pb was, respectively, 44 and 40. Finally, each element was examined in more detail subsequently based on the correlations between Cu and Pb from the least squares regression analysis.

2.2.3. Remote Sensing Data Preparation

Landsat8 images of the study area in 2017 and 2019 were collected from the U.S. Geological Survey (https://earthexplorer.usgs.gov/, accessed on 28 September 2019), and images with cloud interference were excluded. Because there is less vegetation interference from November to March every year, Landsat8 data from this period are more conducive to satellite observation of soil properties. More importantly, the images acquired during this period are closer to the collection time of the soil samples. Therefore, Landsat8 images acquired in December 2017 and in November 2019 were used and then preprocessed for atmosphere correction using the FLAASH module of the ENVI 5.0 software.

In addition, a 30 m digital elevation model (DEM) product was acquired from the geospatial data cloud website (http://www.gscloud.cn/, accessed on 4 January 2020), and then the slope and aspect data were derived using ArcGIS 10.0 software.

3. Methodology

3.1. Optimal Factors of the Metals of Interest

3.1.1. Spectral Factors

Previous studies, e.g., [24], have shown that the spectral curves of heavy-metal-contaminated soil and normal soil showed different spectral characteristics. The heavy-metal-contaminated soil showed strong absorption characteristics in the spectrum range of 400–500 nm in Landsat8 satellite imagery, spectral reflectance showed an overall upward trend from 500 to 780 nm, reflectance showed a downward trend from 780 to 900 nm, and reflectance of polluted soil showed a rising trend from 1200 to 2500 nm. These results indicated that the above four spectrum ranges were diagnostic ranges to distinguish heavy-metal-contaminated soil from normal soil. In this paper, the reflectance of the B2–B7 bands displayed strong correlations with the Pb content of the soils, while those of the B2–B4 bands showed stronger correlations with the Cu content. Therefore, according to the geomorphic types of the study area, the spectrum reflectance of six bands on Landsat8 images B2–B7 was selected as the candidates.

Considering that the heavy metals are often mixed with other soil components and the content of heavy metals contained in soil is usually low, the characteristics of heavy metals in soil are also very weak, especially in satellite imagery; therefore, it is difficult to use the reflectivity or absorption spectrum characteristics of heavy metals to estimate the content of heavy metals in soil. However, the content of soil heavy metals can be indicated indirectly by the adsorption or occurrence relationship among water, clay minerals contained in soil, and environmental factors such as vegetation growth circumstances, topography, and the distance to pollution sources, as referred to by [6,19,25].

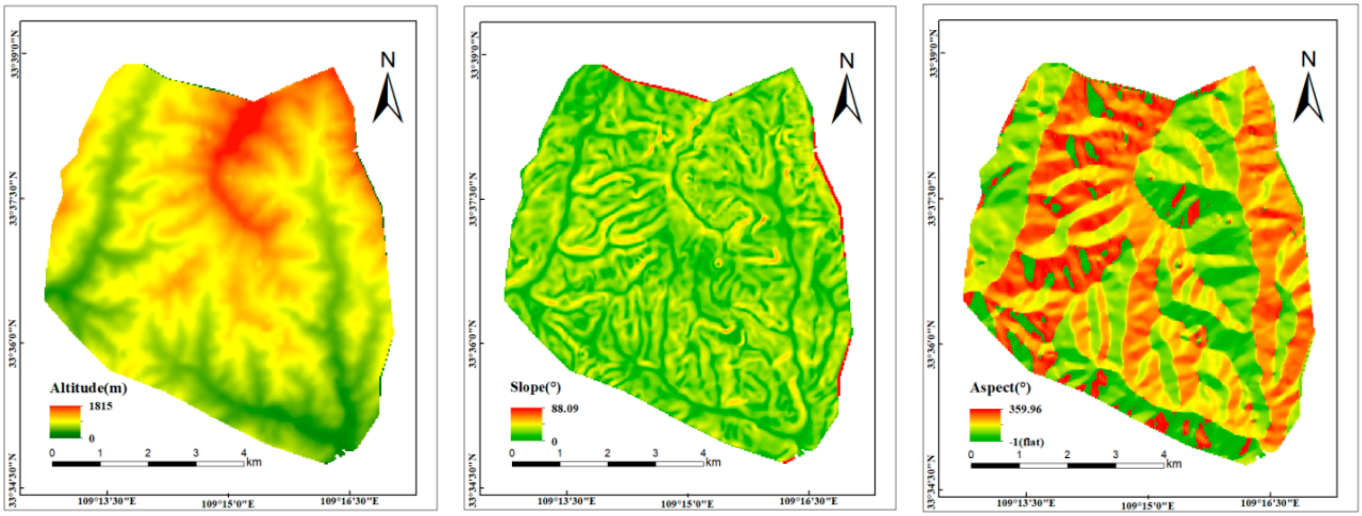
Based on the above analysis, eight spectral indices derived from the spectral reflectance of bands B2–B7 of the Landsat8 image acquired in 2017 and 2019 reflect the soil properties related to heavy metals. Specifically, the clay mineral ratio (CMR) [26] reflects the clay mineral content in soil, which indirectly can affect the distribution of heavy metals in soil. The improved normalized water index (MNDWI) [27] can strengthen the characteristics of soil moisture. In the vegetation coverage area, vegetation growth circumstances reflected by the normalized vegetation index (NDVI), differential vegetation index (DVI), and enhanced vegetation index (EVI) can indirectly reflect the type of soil and content of heavy metals in soil [28]. The greenness, brightness, and humidity components generated by the tasseled cap transformation can discriminate vegetation from soil information; the definition of each spectral index derived from Landsat8 imagery can be seen in Table 1.

Table 1. Spectral indices indicative of the content of Cu and Pb.

Type	Factors	Definition
Spectral index	MNDWI	$(B3 - B6)/(B3 + B6)$
	DVI	$B5/B4$
	CMR	$B6/B7$
	EVI	$2.5 \times (B5 - B4)/(B5 + 6 \times B4 - 7.5 \times B2 + 1)$
	NDVI	$(B5 - B4)/(B5 + B4)$
	Greenness	$-0.294 \times B2 - 0.243 \times B3 - 0.542 \times B4 + 0.728 \times B5 + 0.071 \times B6 - 0.161 \times B7$
	Brightness	$0.303 \times B2 + 0.279 \times B3 + 0.473 \times B4 + 0.56 \times B5 + 0.508 \times B6 + 0.187 \times B7$
	Wetness	$0.151 \times B2 + 0.197 \times B3 + 0.328 \times B4 + 0.341 \times B5 - 0.712 \times B6 - 0.456 \times B7$

3.1.2. Terrain Factors

Previous studies, e.g., [29], have shown that auxiliary factors such as terrain have a great effect on the spatial distribution of heavy metals in soil. Considering that the study area has high mountains and medium mountains partly covered by vegetation, this study introduced three topography factors (altitude, slope, and aspect, as in Figure 3) to describe the spatial distribution of heavy metals in soil.



(**a**) (**b**) (**c**)

Figure 3. Altitude (**a**), slope (**b**), and aspect (**c**) map in the study area.

As shown in Figure 3, the altitude difference in the study area is large and the slope is steep, which makes the heavy metals in soil at the top of the mining area tend to migrate downward. The slope direction will affect the circumstances of vegetation growth and inhibit rain from washing away the heavy metals in soil towards the bottom of the mountain. These auxiliary environmental variables will have a certain impact on the spatial distribution of metals in soil.

3.1.3. Select the Optimal Factors for Each Metal of Interest

To select the optimal factors indicating the content of the two heavy metals, the correlation analysis of six spectral bands, eight spectral indices, and the three topography indicators were made using the least squares method. Subsequently, a collinearity test was performed. According to the detection criteria, the collinearity test between one of the factors and the others is weak if the value of the variance inflation factor (VIF) is less than 10, and the tested factor with high correlation is viewed as one of the optimal indicators. Exceptionally, three terrain factors showed low correlation coefficients; however, the result of a multivariant linear regression with a combination of some terrain factors with

the chosen spectral reflectivity and spectral indices showed an improvement in decision coefficients R^2 and root mean square error (RMSE).

Therefore, aspect and altitude were added to the set of optimal factors for Cu and Pb.

According to the analysis method above mentioned and the previous studies (e.g., [13], the set of optimal spectral factors of Cu and Pb was chosen as in 2017 (Table 2) and in 2019 (Table 3).

Table 2. Correlation coefficients of the optimal factors of Cu and Pb in 2017.

Cu		Pb	
Factors	**Correlation Coefficients**	**Factors**	**Correlation Coefficients**
B2	0.518	B2	0.419
B3	0.466	B3	0.418
B4	0.363	B4	0.428
EVI	−0.364	B6	0.313
CMR	−0.453	B7	0.332
NDVI	−0.371	EVI	−0.326
MNDWI	0.396	Brightness	0.354
Greenness	−0.386	Aspect	−0.262
Aspect	0.023	Elevation	−0.179

Table 3. Correlation coefficients of the optimal factors of Cu and Pb in 2019.

Cu		Pb	
Factors	**Correlation Coefficients**	**Factors**	**Correlation Coefficients**
B2	0.618	B2	0.407
B3	0.598	B3	0.415
B4	0.497	B4	0.401
EVI	−0.372	B6	0.329
CMR	0.516	B7	0.314
NDVI	−0.360	EVI	−0.540
MNDWI	0.447	Brightness	0.365
Greenness	−0.411	Aspect	−0.578
Aspect	−0.569	Altitude	0.415

In addition, this study developed the least square regression analysis method to analyze the correlation among the two metals. The analysis results showed that the correlations between the two metals were greater; thus, the two heavy metals need to be analyzed and estimated separately in the study.

3.2. Model for Soil Heavy Metal Retrieval Using Transfer Learning

3.2.1. Construct a Pre-Trained GA-BP Model Using Samples in 2017

The quantitative retrieval tasks in remote sensing applications usually can be viewed as a statistical regression problem. Generally, the statistical learning or shallow machine learning regression models, such as PLSR, SVR, condition rule-based M5 model tree, extreme learning machine, and random forest, have shown the advantages of low training cost and better performance for a local region.

Compared with others, BP networks are popular for solving complex nonlinear regression problems. The network is characterized by signal forward transmission and an error back propagation structure. The network weights are dynamically adjusted with the estimation error by back propagation during gradient descent. However, the method that randomly initializes the weights and thresholds of the original BP network often leads to local optimization [30]. Although the distributionally robust optimization (DRO) algorithm [31] was proposed for different applications, e.g., network behavior analysis and risk management, the genetic algorithm (GA) is popularly used to seek global optimization for nonlinear problems.

In our previous work [14], the BP network optimized by GA was compared to multivariate linear regression model and M5 model tree for the retrieval of soil heavy metal in the study area in 2017. It has shown that our selected GA-BP approach perform well. Therefore, this study introduced GA to initialize the weights and thresholds of a three-layer BP network by the optimal individual selection to improve the accuracy and stability of the approximation.

The main steps of the GA-BP model are as follows:(1) determine the structure of the BP network; (2) initialize the GA population and train the BP network with training samples; (3) train the GA-BP network. The parameters of the GA-BP network are listed in Table 4.

Table 4. The setting of the GA-BP network structure and parameters.

Number of Input Layer Neurons	**Number of Hidden Layer Neurons**	**Number of Output Layer Neurons**	**Weight Joining Input Layer with Hidden Layer**	**Threshold between Input and Hidden Layer**
9	4	1	40	5
Number of population	Maximum of evolutionary		Crossover probability	Mutation probability
30	50		0.3	0.1
Optimization algorithm	Maximum of iterations		training accuracy	learning rate
Levenberg Marquardt	50		0.3	0.1

Thus, a GA-BP network was established as suitable for the estimation of the content of Cu and Pb in which 80% of the randomly selected soil samples acquired in 2017 were selected to train the weight parameters of the above GA-BP model.

3.2.2. Construct Our Tr-GA-BP Model for Retrieval of Heavy Metals in 2019

To reduce the soil sampling costs for heavy metals in 2019 and to avoid a negative transfer from source domain to target domain, which is adverse to improving the estimation precision of soil heavy metal content in the study area in 2019, the study proposed the Tr-GA-BP model using a parameter transfer learning strategy based on the pre-trained GA-BP network. The idea of the proposed Tr-GA-BP model is that the optimal individuals of the GA-BP network were transferred to the domain in 2019 through similarity analysis between the feature from source domain(referring to the study area in 2017) and the target domain(referring to the study area in 2019). Consequently, the parameters of the GA-BP model were retrained using a few samples collected in 2019.Here, the gradient descent method was used to optimize the parameters of the pre-trained GA-BP model in the process of the similarity analysis on the features between the source domain and the target domain. The description of our proposed Tr-GA-BP model was described as follows:

Let $D_S = \left\{ \left(X_S^1, Y_S^1\right), ..., \left(X_S^i, Y_S^i\right), \ldots, \left(X_S^M, Y_S^M\right) \right\}$ denote a set of samples from the area in 2017, and $\left(X_S^i, Y_S^i\right)$ represents the i-th sample (M is the number of samples from source domain). $X_S^i \in R^L$ represents the L-dimension feature vector defined as the optimal factors of the source domain sample and $Y_S^i \in R$ is the one-dimensional vector representing the measured content of Cu and Pb contained in the soil samples acquired in 2017.

Let $D_T = \left\{ \left(X_T^1, Y_T^1\right), ..., \left(X_T^j, Y_T^j\right), \ldots, \left(X_T^N, Y_T^N\right) \right\}$ denote a set of samples from the same area in 2019, and $\left(X_T^j, Y_T^j\right)$ represents the j-th sample (N is the number of samples from the target domain). $X_T^j \in R^L$ represents the L-dimension consisting of the optimal factors of the target domain and $Y_T^j \in R$ is the one-dimensional vector representing the measured content of Cu and Pb contained in the soil samples acquired in 2019.

Let W_S^* designate the optimal parameter matrix learned from the pre-trained GA-BP model, respectively. A similarity coefficient β is defined to measure the similarity of parameters between the source domain and the target domain. Let W_T be a parameter matrix of the Tr-GA-BP network in the target domain. It can be defined as Equation (1):

$$W_T = \beta W_s^* \tag{1}$$

where the initial value β^0 of β is obtained using the grid search algorithm with the range of [0, 0.1, 1]. Then, fewer samples from the target domain were used to retrain the parameters of the pre-trained GA-BP network. Thus, the matrix W_T was updated with the similarity coefficient β optimized using the gradient decent algorithm. Finally, the parameters β and W_T of our Tr-GA-BP model could reach the optimum simultaneously.

The construction steps of the Tr-GA-BP model can be described as follows:

(1) Obtain the optimal parameters matrix W_S^* (including weight and threshold parameters) using the pre-trained GA-BP model from the source domain.

(2) Set the initial value of similarity coefficient β as β^0 with a range of [0,0.1,1] using the grid search algorithm and initialize the parameter matrix W_T of target domain as $W_T^0 = \beta^0 W_s^*$.

(3) Retrain the GA-BP model using a few samples from target domain using Equation (1) to update the optimal similarity coefficient β^*; then, the optimal parameters matrix W_T^* is obtained.

Thus, the Tr-GA-BP model can be formed based on the transfer learning idea for the retrieval of Cu and Pb content in 2019 in the study area.

According to the proposed Tr-GA-BP model, let the initial value of the similarity coefficient between the source domain and the target domain be $\beta^0 = 0.1$, and the optimal weight matrix $W_{T,Cu}^*$ for Cu and $W_{T,Pb}^*$ for Pb, respectively, are obtained using samples acquired in 2017 and 2019:

$$W_{T,Cu}^* = \begin{bmatrix} 0.9912 & 0.3466 & 0.6298 & 1.7965 \\ 2.2955 & 1.6426 & 0.4515 & -0.2872 \\ 0.7061 & 2.0578 & 0.5059 & 2.4378 \\ -5.9234 & -0.8229 & -0.0051 & 0.0037 \\ 0.0032 & 0.4729 & 2.0399 & -2.2633 \\ 0.0494 & 4.4254 & 2.8500 & -2.1472 \\ 2.8734 & -1.3132 & -3.9708 & 0.8576 \\ 0.9380 & -1.3607 & -1.9516 & -2.1785 \\ 3.3058 & -2.7859 & -0.7225 & 1.3213 \end{bmatrix}, W_{T,Pb}^* = \begin{bmatrix} 0.9802 & 0.2591 & 0.5308 & 1.2344 \\ 1.3934 & -1.5672 & 1.2905 & -2.0218 \\ 2.0072 & 1.8290 & 1.3902 & 2.1033 \\ -4.0457 & -1.0023 & -0.8964 & 0.0109 \\ 0.1053 & 0.5835 & 1.9021 & -1.9234 \\ 0.0905 & 4.6349 & -2.6491 & 1.9202 \\ 1.2335 & -1.9742 & -2.6356 & 1.0923 \\ 1.0589 & -1.5800 & -1.5588 & -2.3487 \\ 1.9321 & -2.0671 & 0.9522 & -1.3568 \end{bmatrix}$$

The optimal threshold matrix $T_{T,Cu}^*$ and $T_{T,Pb}^*$ for Cu and Pb, respectively, is:

$$T_{T,Cu}^* = \begin{bmatrix} -4.355 \\ -7.244 \\ -7.698 \\ -2.663 \\ -0.631 \\ -0.768 \\ -0.894 \\ -1.320 \\ -0.932 \end{bmatrix}, \; T_{T,Pb}^* = \begin{bmatrix} -2.190 \\ -2.923 \\ -5.992 \\ -3.489 \\ -1.578 \\ -0.904 \\ -1.902 \\ -0.369 \\ -1.790 \end{bmatrix}$$

4. Implementation and Results

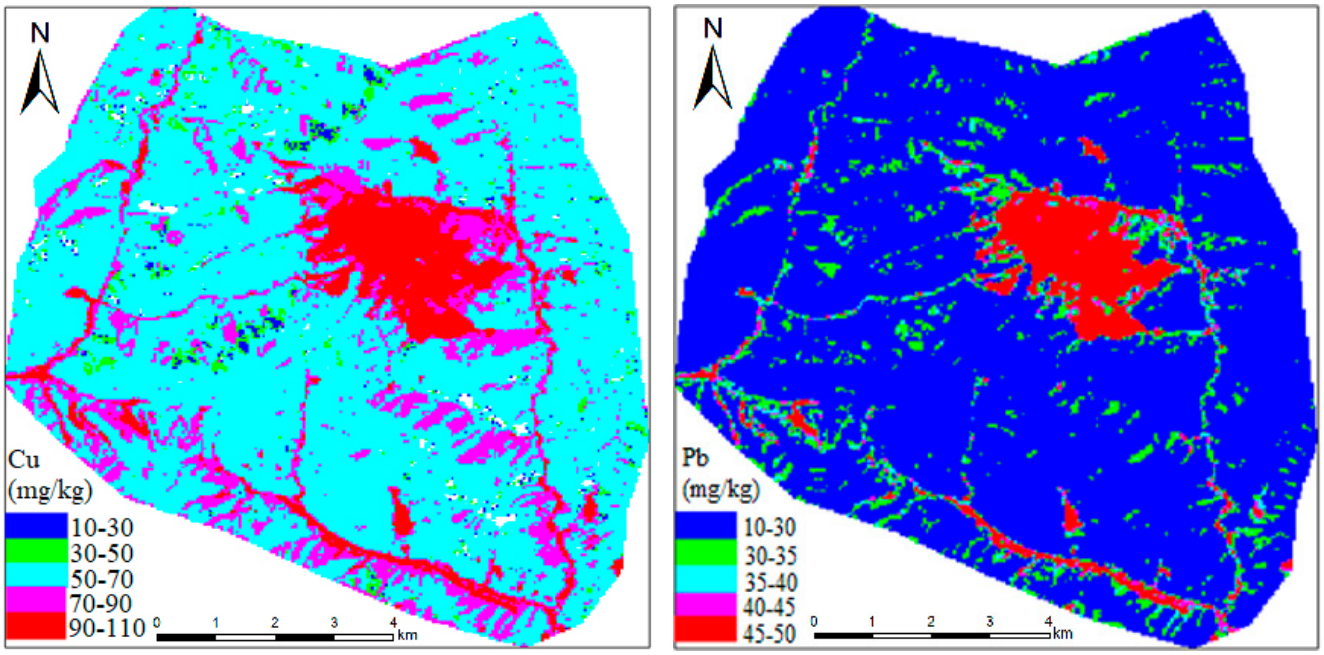
4.1. Implementation

As stated in the former section, we obtain a larger set of samples, D_S, for Cu and Pb in soil from the source domain in 2017 and a smaller set of samples, D_T, from the target domain in 2019. Subsequently, the GA-BP model was pre-trained using 44 samples collected in the

study area in 2017. Furthermore, the proposed Tr-GA-BP model was formed by retraining GA-BP model with only 16 samples from the target domain in 2019.The implementation of estimating soil heavy metal content in the study area in 2019 using our well-trained Tr-GA-BP model was performed under Windows using MATLAB programming language

4.2. Results and Discussion

The content of Cu and Pb at each site in the study area in 2019 was estimated using our Tr-GA-BP model mentioned above; the spatial distribution of the estimated content of both elements was mapped as shown in Figure 4.



(**a**) (**b**)

Figure 4. Spatial distribution map of Cu (**a**) and Pb (**b**) content in 2019 using our model.

As seen from Figure 4, the higher content of Cu and Pb in 2019 mainly was present in the mining area, slag stacking area, and on both sides of the road in the study area. According to the field survey, the ore is always transported from the mining area at the top of the slope to the road in the valley. Due to the accumulation of fallen ore, the metal content of the road is high. Therefore, the spatial distribution of the estimated results of Cu and Pb in the area using our proposed Tr-GA-BP model are consistent with the field validation and the result from our previous study using a different method [13].

To quantitatively verify the effectiveness of the proposed Tr-GA-BP model in this paper, the remaining 20% of samples was used to evaluate the estimation error for the content of Cu and Pb, taking RMSE and mean relative error (MRE) as measures, as in Table 5.

Table 5. The estimation error of our models for Cu and Pb in 2019.

	Cu		**Pb**	
	GA-BP Model	**Tr-GA-BP Model**	**GA-BP Model**	**Tr-GA-BP Model**
RMSE	13.432	9.078	4.390	2.804
MRE	1.902	0.369	1.753	0.521

From Table 5, RMSE and MRE values using the proposed Tr-GA-BP model were, respectively, 9.078 and 0.369 for Cu, reduced by 4.688 and 1.533 compared to the GA-BP

model. For Pb, the RMSE and MRE values using the proposed Tr-GA-BP model were 2.804 and 0.521, respectively, which reduced by 1.586 and 1.232 compared to the GA-BP model. Thus, it was proved that the accuracy of our proposed Tr-GA-BP model based on transfer learning and prior information is effective in improving the precision of soil heavy metal content estimation in the case of fewer samples from target domain and is superior to that of the traditional GA-BP network for the estimation error of Cu and Pb content in topsoil.

To explore the degree of soil contaminated by Cu and Pb, a comparison of the content of Cu and Pb in 2019 with the reference value (i.e., the maximum and arithmetic mean values of soil element content) published by the Shaanxi Province in China is listed in Table 6.

Table 6. Contrast of the estimated content and the reference value for both elements(unit: mg/kg).

	Minimum		**Maximum**		**Average**		**Standard Derivation**	
	Estimated	**Reference**	**Estimated**	**Reference**	**Estimated**	**Reference**	**Estimated**	**Reference**
Cu	23	6.80	82	43.60	49.37	21.40	15.81	7.74
Pb	8.1	13.70	38.9	34.50	21.45	21.40	6.88	5.04

From Table 6, it was found that both the maximum and the arithmetic average of Cu content in 2019 estimated here are far greater than the corresponding reference value. For Pb, the estimated value is slightly greater than its corresponding reference value and the average is close to the reference value.

To further investigate the circumstance of the spatial distribution of both elements, a statistical analysis of the estimated content of Cu and Pb is listed as shown in Table 7.

Table 7. Statistical results of the estimated content of Cu and Pb in 2019.

Cu	**Content(mg/kg)**	0–30	30–50	50–70	70–90	90–110
	Percent (%)	2.6	2.4	80.1	6.9	8.4
Pb	**Content(mg/kg)**	0–30	30–35	35–40	40–45	45–50
	Percent (%)	84.1	4.9	1.6	2.1	7.3

From Figure 4 and Tables 6 and 7, conclusions can be drawn: the estimated value of the Cu content changes in the range from 0 to 110 mg/kg and the Cu content ranges from 50 to 70 mg/kg, accounting for 80.1% of the total study area in 2019. The Pb content in the area ranges 10–50 mg/kg. The area with the content no more than 30 mg/kg accounted for 84.1% of the total area. In addition, the content of Cu is the highest; Pb is the second highest in the study area, which is consistent with the geochemical investigation mapping data [21]. Meanwhile, by comparing the two elements estimated in this study with the maximum and arithmetic mean values of the reference values of soil elements in Shaanxi Province, it is found that the content of the two elements in some parts of the study area exceed the average reference values of soil elements in Shaanxi Province, which indicates that the soil in some areas has been polluted by these two heavy metals since 1990.

5. Conclusions

Retrieving the content of topsoil heavy metals at a lower sample collection cost for environmental monitoring in a mining area while keeping high estimation precision is challenging. Considering the Daxigou mining area in Shaanxi located in the Qinling Mountains and covered by vegetation as a study case, this study introduced the transfer learning idea to innovatively construct a Tr-GA-BP network so as to implement the retrieval of the content of two interesting heavy metals, i.e., Cu and Pb, in soil in 2019 based on a pre-trained GA-BP network using Landsat8 multispectral satellite images, DEM, and geochemical data using more samples collected in 2017 and less samples in 2019.

Finally, the spatial distribution mapping and content change analysis were conducted using the proposed Tr-GA-BP network. From the validation results using field-collected data, the RMSE and MRE values using the proposed Tr-GA-BP model were, respectively, 9.078 mg/kg and 0.369 mg/kg for Cu, reduced by 4.688 mg/kg and 1.533 mg/kg compared to the GA-BP model. For Pb, the RMSE and MRE values using the proposed Tr-GA-BP model were 2.804 mg/kg and 0.521 mg/kg, respectively, which reduced by 1.586 mg/kg and 1.232 mg/kg compared to the GA-BP model. Thus, our proposed Tr-GA-BP model based on transfer learning and prior information performs well in improving the estimation precision of Cu and Pb content in soil under the condition of16training samples collected in 2019 and is superior to that of the traditional GA-BP network. In addition, the content of Cu is the highest; Pb is the second highest in the study area. Both of them were mainly distributed in the exploitation, slag stacking, on the roadsides, and at the base of slope, which is consistent with the field investigation results and our previous study result with different methods. This pollution has been endangering the soil, water, and the health of local residents.

The proposed method in this paper should show better performance if more soil samples are collected. In the future, a transfer learning strategy should be optimized and terrain illumination and shadow effects in mountainous areas should be considered so as to further improve the estimation accuracy of the heavy metal content in soil.

Author Contributions: Conceptualization, Y.Y., R.C. and A.H.; Methodology, Y.Y., Q.C. and R.C.; Software, Q.C.; Validation, Q.C.; Formal analysis, Y.Y., R.C., A.H. and Y.W.; Investigation, Y.Y. and Q.C.; Resources, Y.Y. and Y.W.; Data curation, Y.Y.; Writing—original draft, Q.C.; Writing—review & editing, Y.Y., R.C., A.H. and Y.W.; Supervision, Y.Y.; Project administration, Y.Y.; Funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: The study was jointly supported by the Natural Science Basis Research Plan in Shaanxi Province of China (No.2022JM-163), the Basic Scientific Research Business of Central University of Chang'an University (No.300102269205), and National Natural Science Foundation of China (No.42261144749 and 41877232).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data is unavailable due to privacy.

Acknowledgments: The authors thank the reviewer and editors for their constructive and helpful comments. The authors are thankful to NASA Land Processes Distributed Active Archive Center User Services, USGS Earth Resources Observation and Science (EROS) Center for providing the Landsat8 data, and the Geospatial Data Cloud in China for providing DEM data.

Conflicts of Interest: The authors declared that there are no financial or nonfinancial conflict of interest that are directly or indirectly related to this work.

References

1. Wu, Y.; Li, X.; Yu, L.; Wang, T.; Wang, J.; Liu, T. Review of soil heavy metal pollution in China: Spatial distribution, primary sources, and remediation alternatives. *Resour. Conserv. Recycl.* **2022**, *181*, 106261. [CrossRef]
2. Krasowska, M.; Kowczyk-Sadowy, M.; Szatyłowicz, E.; Obidziński, S. Analysis of the Possibility of Heavy Metal Ions Removal from Aqueous Solutions on Fruit Pomace. *J. Ecol. Eng.* **2023**, *24*, 169–177. [CrossRef]
3. Łapiński, D.; Wiater, J.; Szatyłowicz, E. The Content of Heavy Metals in Waste as an Indicator Determining the Possibilities of their Agricultural Use. *J. Ecol. Eng.* **2019**, *20*, 225–230. [CrossRef]
4. Huo, A.; Wang, X.; Zhao, Z.; Yang, L.; Zhong, F.; Zheng, C.; Gao, N. Risk Assessment of Heavy Metal Pollution in Farmland Soils at the Northern Foot of the Qinling Mountains. *Int. J. Environ. Res. Public Health* **2022**, *19*, 14962. [CrossRef]
5. Tong, W.; Liu, J.; Fei, L.; Sun, Z. Inversion of soil heavy metals in Guanzhong area of Shaanxi based on VIS-NIR spectroscopy. *J. Phys. Conf. Ser.* **2020**, *1549*, 022145. [CrossRef]
6. Xu, X.; Chen, S.; Ren, L.; Han, C.; Lv, D.; Zhang, Y.; Ai, F. Estimation of Heavy Metals in Agricultural Soils Using Vis-NIR Spectroscopy with Fractional-Order Derivative and Generalized Regression Neural Network. *Remote Sens.* **2021**, *13*, 2718. [CrossRef]

7. Gu, Y.; Zhao, N.; Ma, M.; Meng, D.S.; Jia, Y.; Fang, L.; Liu, J.-G.; Liu, W.-Q. Mapping Analysis of Heavy Metal Elements in Polluted Soils Using Laser-Induced Breakdown Spectroscopy. *Spectrosc. Spectr. Anal.* **2018**, *38*, 982–989.
8. Tan, K.; Wang, H.; Chen, L.; Du, Q.; Du, P.; Pan, 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]
9. Zhang, Y.; Xu, Y.; Xiong, W.; Qu, R.; Ten, J.; Lou, Q.; Lv, N. Inversion Study of Heavy Metals in Soils of Potentially Polluted Sites Based on UAV Hyperspectral Data and Machine Learning Algorithms. In Proceedings of the 2021 11th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing, Amsterdam, The Netherlands, 24–26 March 2021.
10. Wang, F.; Gao, J.; Zha, Y. Hyperspectral sensing of heavy metals in soil and vegetation: Feasibility and challenges. *ISPRS J. Photogrammtry Remote Sens.* **2018**, *136*, 73–84. [CrossRef]
11. Liu, Y.; Luo, Q.; Cheng, H. Application and development of hyperspectral remote sensing technology to determine the heavy metal content in soil. *J. Agro-Environ. Sci.* **2020**, *39*, 2699–2709.
12. Peng, Y.; Kheir, R.B.; Adhikari, K.; Malinowski, R.; Greve, M.B.; Knadel, M.; Greve, M.H. Digital Mapping of Toxic Metals in Qatari Soils Using Remote Sensing and Ancillary Data. *Remote Sens.* **2016**, *8*, 1003. [CrossRef]
13. Wang, T.; Zhao, M.; Yang, Y.; Zhang, Y.; Cui, Q.-F.; Li, L.-T. Inversion of Heavy Metals Content in Soil Using Multispectral Remote Sensing Imagery in Daxigou Mining Area of Shaanxi. *Spectrosc. Spectr. Anal.* **2019**, *39*, 3880.
14. Cui, Q. *Research on Remote Sensing Estimation and Monitoring Method of Heavy Metal Content in Mining Area Soil Based on Machine Learning*; Chang'an University: Xi'an, China, 2020.
15. Zhou, W.; Yang, H.; Xie, L.; Li, H.; Huang, L.; Zhao, Y.; Yue, T. Hyperspectral inversion of soil heavy metals in Three-River Source Region based on random forest model. *Catena* **2021**, *202*, 105222. [CrossRef]
16. Xia, X.; Yang, J. Molecular sequestration mechanisms of heavy metals by iron oxides in soils using synchrotron- based techniques: A review. *J. Appl. Ecol.* **2019**, *30*, 348–358.
17. Wang, H.; Tan, K.; Wu, F.; Chen, Y.; Chen, L.-H. Study of the Retrieval and Adsorption Mechanism of Soil Heavy Metals Based on Spectral Absorption Characteristics. *Spectrosc. Spectr. Anal.* **2020**, *40*, 316–323.
18. Chen, W.; Peng, L.; Hu, K.; Zhang, Z.; Peng, C.; Teng, C.; Zhou, K. Spectroscopic response of soil organic matter in mining area to Pb/Cd heavy metal interaction: A mirror of coherent structural variation. *J. Hazard. Mater.* **2020**, *393*, 122425. [CrossRef]
19. 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]
20. Tao, C.; Wang, Y.; Zou, B.; Tu, Y.L.; Jiang, X.L. Assessment and Analysis of Migrations of Heavy Metal Lead and Zinc in Soil with Hyperspectral Inversion Model. *Spectrosc. Spectr. Anal.* **2018**, *38*, 1850–1855.
21. Fang, W.X.; Hub, R.Z.; Huang, Z.Y. Mineralization Zoning in Yindongzi-Daxigou Barite-Siderite, Silver-Polymetallic Deposits in the Qinling Orogen. *Chin. J. Geochem.* **2001**, *20*, 45–51. [CrossRef]
22. Chen, L.; Wang, J.; Gu, T. Study on Fuzzy Comprehensive Evaluation of Ecological Environment in Daxigou Iron Mine. *Chin. J. Soil Sci.* **2017**, *48*, 794–799.
23. State Environmental Protection Administration (Ed.) *Background Value of Soil Elements in China*; China Environmental Science Press: Beijing, China, 1990; pp. 10–13.
24. Jin, J.; Zhou, X.P.; Ma, K. The analysis of the optimum band combination based on the image of Landsat-8—A case study of thematic survey of heavy metals in soils in the mining area of Bayan Obo. *Inn. Mong. Univ. Sci. Technol.* **2016**, *35*, 37–41.
25. Zhang, C.; Ren, H.; Dai, X.; Qin, Q.; Li, J.; Zhang, T.; Sun, Y. Spectral characteristics of copper-stressed vegetation leaves and further understanding of the copper stress vegetation index. *Int. J. Remote Sens.* **2019**, *40*, 4473–4488. [CrossRef]
26. Hakan, M. Mineral composite assessment of Kelkit River Basin in Turkey by means of remote sensing. *J. Earth Syst. Sci.* **2009**, *118*, 701–710.
27. Xu, H.Q. Modification of Normalized Difference Water Index (NDWI) to Enhance Open Water Features in Remotely Sensed Imager. *Int. J. Remote Sens.* **2006**, *27*, 3025–3033. [CrossRef]
28. Piekarczyk, J.; Kazmierowski, C.; Krolewicz, S. Relationships between soil properties of the abandoned fields and spectral data derived from the advanced spaceborne thermal emission and reflection radiometer (ASTER). *Adv. Space Res.* **2012**, *49*, 280–291. [CrossRef]
29. Kheir, R.B.; Shomar, B.; Greve, M.B.; Greve, M.H. On the quantitative relationships between environmental parameters and heavy metals pollution in Mediterranean soils using GIS regression-trees: The case study of Lebanon. *J. Geochem. Explor.* **2014**, *147*, 250–259. [CrossRef]
30. Zhang, S.; Wang, Z.; Zou, X.; Qian, Y.; Yu, L. Recognition of tea disease spot based on hyperspectral image and genetic optimization neural network. *Trans. Chin. Soc. Agric. Eng.* **2017**, *33*, 200–207.
31. Li, B.; Tan, Y.; Wu, A.; Duan, G. A distributionally robust optimization based method for stochastic model predictive control. *IEEE Trans. Autom. Control* **2021**, *67*, 5762–5776. [CrossRef]