water

MDPI

Article

Time-Domain Transfer Learning for Accurate Heavy Metal Concentration Retrieval Using Remote Sensing and TrAdaBoost Algorithm: A Case Study of Daxigou, China

Yun Yang [1,2,*], Qingzhen Tian [1], Han Bai [3], Yongqiang Wei [4,5], Yi Yan [1] and Aidi Huo [6,*]

[1] College of Geological Engineering and Geomatics, Chang'an University, Xi'an 710054, China
[2] Key Laboratory of Formation Mechanism and Prevention and Control of Mine Geological Disasters, Ministry of Natural Resources, Xi'an 710054, China
[3] State Grid Location Based Service Co., Ltd., Beijing 102209, China
[4] Xi'an Research Institute of Surveying and Mapping, Xi'an 710054, China
[5] State Key Laboratory of Geo-Information Engineering, Xi'an 710054, China
[6] School of Water and Environment, Chang'an University, Xi'an 710054, China
* Correspondence: yangyunbox@chd.edu.cn (Y.Y.); huoaidi@chd.edu.cn (A.H.)

Abstract: Traditionally, the assessment of heavy metal concentrations using remote sensing technology is sample-intensive, with expensive model development. Using a mining area case study of Daxigou, China, we propose a cross-time-domain transfer learning model to monitor heavy metal pollution using samples collected from different time domains. Specifically, spectral indices derived from Landsat 8 multispectral images, terrain, and other auxiliary data correlative to soil heavy metals were prepared. A cross time-domain sample transfer learning model proposed in the paper based on the TrAdaBoost algorithm was used for the Cu content mapping in the topsoil by selective use of soil samples acquired in 2017 and 2019. We found that the proposed model accurately estimated the concentration of Cu in the topsoil of the mining area in 2019 and performed better than the traditional TrAdaBoost algorithms. The goodness of fit (R^2) of the test set increased from 0.55 to 0.66; the relative prediction deviation (RPD) increased from 1.37 to 1.76; and finally, the root-mean-square deviation (RMSE), decreased from 8.33 to 7.24 mg·kg^{-1}. The proposed model is potentially applicable to more accurate and inexpensive monitoring of heavy metals, facilitating remediation-related efforts.

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

Citation: Yang, Y.; Tian, Q.; Bai, H.; Wei, Y.; Yan, Y.; Huo, A. Time-Domain Transfer Learning for Accurate Heavy Metal Concentration Retrieval Using Remote Sensing and TrAdaBoost Algorithm: A Case Study of Daxigou, China. *Water* **2024**, *16*, 1439. https://doi.org/10.3390/w16101439

Academic Editor: Laura Bulgariu

Received: 25 March 2024
Revised: 28 April 2024
Accepted: 6 May 2024
Published: 17 May 2024

1. Introduction

According to previous investigations and studies, e.g., [1–3], land worldwide, such as in the United States, Europe and China, is contaminated with heavy metals and other toxins to varying degrees. Heavy metals are the primary source of pollution in mining areas or around smelting plants in China, with land degradation and ecological and human health impacts [2,3]. The efficient monitoring of heavy metal pollution is important, as it facilitates efforts to either prevent or remediate heavy metal pollution. Ground hyperspectral technology allows for the acquisition of soil-linked continuous spectral information, yielding accurate estimates of soil organic matter, heavy metals, and other soil components; however, this method cannot be used for large-scale monitoring. Hyperspectral remote sensing imaging can be used to conduct large-scale soil parameter investigations and monitoring. For example, Han et al. [4] used domestic GF-5 hyperspectral satellite imagery to retrieve the content of heavy metals in the soil based on the XgBoost algorithm, and Zhang et al. [5] used GF-5 hyperspectral remote sensing images to develop an inversion model intended to estimate the content of heavy metals in soil. However, the previous inversion methods using hyperspectral technology have shown limitations and challenges as stated in Ref. [6]. The acquisition cost of commercial hyperspectral imagery with a high spectral and spatial

resolution, such as Zhuhai-1 satellite-based, airborne imaging platforms, etc., is currently relatively high, despite being more favourable for detecting heavy metal concentrations in the soil.

To reduce the cost of remote sensing data acquisition, efforts have been made to retrieve heavy metal concentrations in soil. This is achieved using multispectral remote sensing images in conjunction with auxiliary data that indicate the spatial distribution of heavy metals in soil rather than using expensive hyperspectral remote sensing imagery. For example, Bou et al. [7] combined multiple environmental variables to develop an inversion model for the estimation of soil heavy metal content and verified its effectiveness. Ghrefat et al. [8] proposed integrating remote sensing data, geochemical, GIS and statistical data to assess heavy metal contamination in soil. Yu et al. [9] estimated the nitrogen content in wheat using UAV-based and satellite multispectral imagery, together with topographic, plant and soil metrics using Random Forests (RF) and support vector regression (SVR) models. Several studies (e.g., [10,11]) have also demonstrated that an alternative to the retrieval of heavy metal concentrations in soil is using low-cost multispectral imaging supplemented with multiple environmental variables instead of high-cost hyperspectral images.

In view of the methodology adopted in the retrieval of heavy metal concentrations in soil, most of the abovementioned inversion models were developed based on the partial least squares regression model of the linear hypothesis. In recent years, many machine learning models, such as random forests, have been widely used, such as by [12–14]. Compared to the currently popular deep learning networks, these machine learning models have a simple structure, are rapidly trained, and can be applied toward modelling nonlinear systems. For example, the performance of AdaBoost, owing to a weighted random forest combination, is generally better than a traditional random forest and support vector machine with equal weight, and its training efficiency is high when evaluated through the lens of regression accuracy [12].

Previous research has contributed to characterising heavy metal concentrations in the topsoil (i.e., through various approaches), however, traditional machine learning models exhibit weak generalizability when applied to soil heavy metal content inversion tasks, even across different spatiotemporal contexts. This results in pollution-related investigations and long-term monitoring tasks that are plagued by low operation efficiency and high sample cost. Therefore, improving the accuracy of soil heavy metal content inversion within the context of using few samples is a global challenge.

Transfer learning theory is a solution to the problem of parameter learning when the data distribution between source and target domains differs, as in the case of a few training samples [15]. Remote sensing studies, e.g., Zhang et al. [16], have mainly focused on the extraction of qualitative information, such as object recognition. Recent studies have explored the transferability of soil heavy metal content inversion models and training samples from one region to another [17]. Furthermore, Liu et al. [18] constructed a transfer learning model based on a convolutional neural network using LUCAS open soil spectral dataset and verified the effectiveness of the model for recording clay concentrations using a few samples.

Collecting a substantial number of training samples for soil contamination monitoring is expensive, particularly in mountainous terrain or woodland environment. Moreover, collecting training data for the parameters of machine learning models is expensive, unlike in the field of target detection and classification. Fortunately, transfer learning theory has emerged as an alternative solution in the application of soil heavy metals retrieval; it can reduce the data acquisition cost by transferring sample data or other useful information from one task to another. For example, Tao et al. [19] used the transfer component analysis method to reduce the differences in the probability distribution of soil samples from two or more regions, enabling the application of a model developed using samples from one region to other regions. Wang et al. [20] adopted the transfer component analysis method to analyse the distribution differences between artificial and natural samples

to improve the regression accuracy under limited heavy metal samples data using an expanded training dataset.

We have also made efforts to estimate the content of heavy metals based on transferring learning theory using remote sensing technology; however, a key issue on how to alleviate negative transfers of training samples, which keep higher regression precision, must be solved. Therefore, we proposed a novel inversion model for the retrieval of heavy metal (e.g., element Cu) concentrations in topsoil based on a traditional TrAdaBoost algorithm. This model should allow the accurate and efficient characterization of heavy metal concentrations with less field-measured data by integrating the Landsat 8 image spectrum and its transform spectrum, as well as a variety of auxiliary data. Reduced costs and higher efficiency in characterising soil heavy metal contamination may facilitate more robust efforts to prevent soil heavy metal pollution. Furthermore, our findings provide information for decision-making within the context of land degradation and ecological management, and restoration in mining areas.

2. Materials and Methods

2.1. Materials

2.1.1. Overview of the Study Area

The Daxigou mining area, approximately 4.33 km^2 in Zhashui County, Shangluo City of Shaanxi province in China, in the hinterlands of the Qinling Mountains, at an altitude of 1000–1600 m (Figure 1), is the case study in the paper.

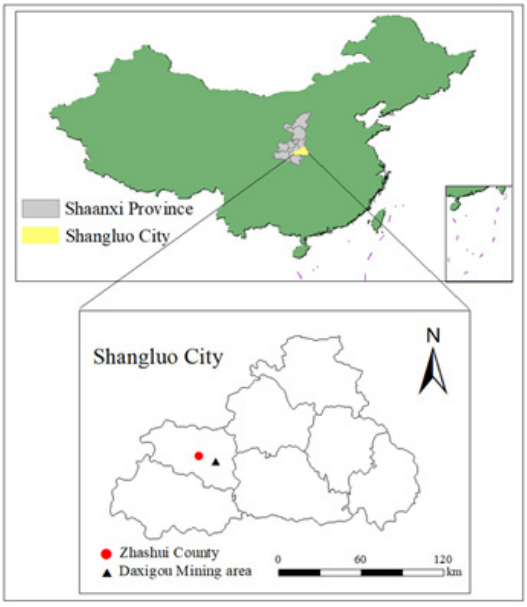


Figure 1. A map of the study area.

The mine is one of the largest siderite mines in China, accounting for 47.6% of the total iron ore reserves in Shaanxi Province. In 1982, it was discovered that the Daxigou-Yindongzi deposit was rich in Cu, Pb, Zn, Ag, and other minerals [21]. The mining activity officially began through open-pit mining in 1988. The long-term mining activities have severely impacted the land with heavy metal pollution, with the loss of ecological integrity [22]. These changes warrant continuous monitoring of the mining area and its surroundings for heavy metal pollution and the development of scientifically informed strategies to address pollution effects.

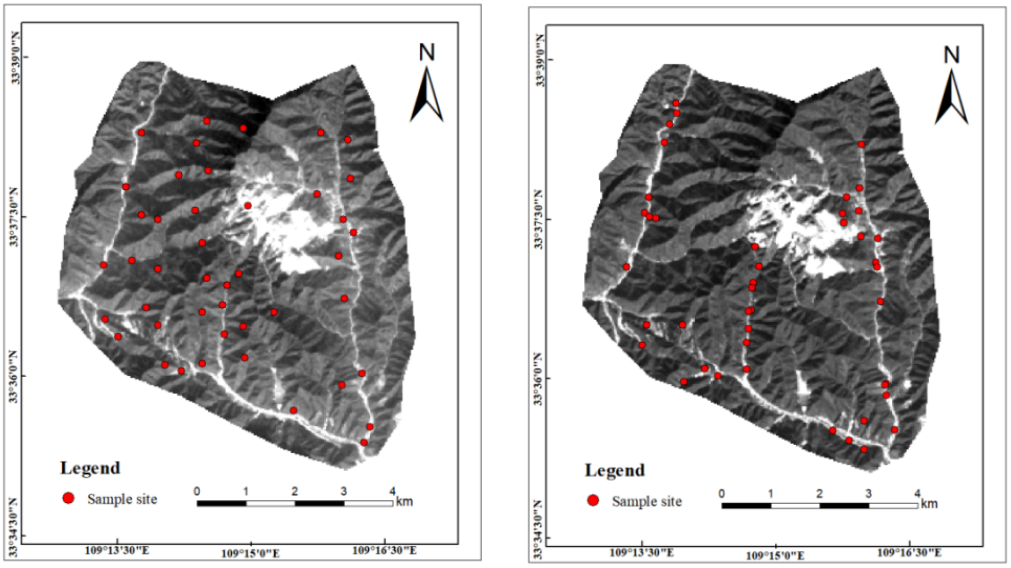
The study area has a complex terrain with medium and low gullies, tectonically eroded landforms, slag piles that are stacked in the valleys and waste dumps in the west and north of the slope, which have had a great impact on the landform and landscape.Land use types include mining areas, cultivated land, forestland, grasslands, industrial and mining

facilities, and residential areas. Of these, the mining areas, industrial and mining facilities and forestland account for most of the land area. Other categories cover a small portion. The soils in the area have relatively high copper–lead concentrations.

2.1.2. Data Preparation

Heavy Metals Data

Soil samples were collected along the three main ridge lines, based on the characteristics of terrain and land cover types of the study area, to ensure that they were fully representative of the area. Most sampling points were in the middle of mountain slopes and near valleys. Heavy metals accumulated in the valley soils due to erosion with a sampling depth of approximately 20–30 cm, whereas the sampling depth in the middle of the slope was approximately 10–20 cm. After sampling, an area of 30 m × 30 m was isolated, and 1 kg of soil was extracted and placed in a dedicated sample bag. The WGS84 coordinates of the centre point of the sampling area for each sample were also recorded, as were the soil properties and environmental parameters. Professional technicians, from the laboratory of environmental testing center of Guolian Quality Inspection Technology Co., Ltd. in Xi'an, China, collected 44 and 43 soil samples using professional instruments from the area in October 2017 and 2019, respectively. The distribution of the sampling sites is shown in Figure 2.



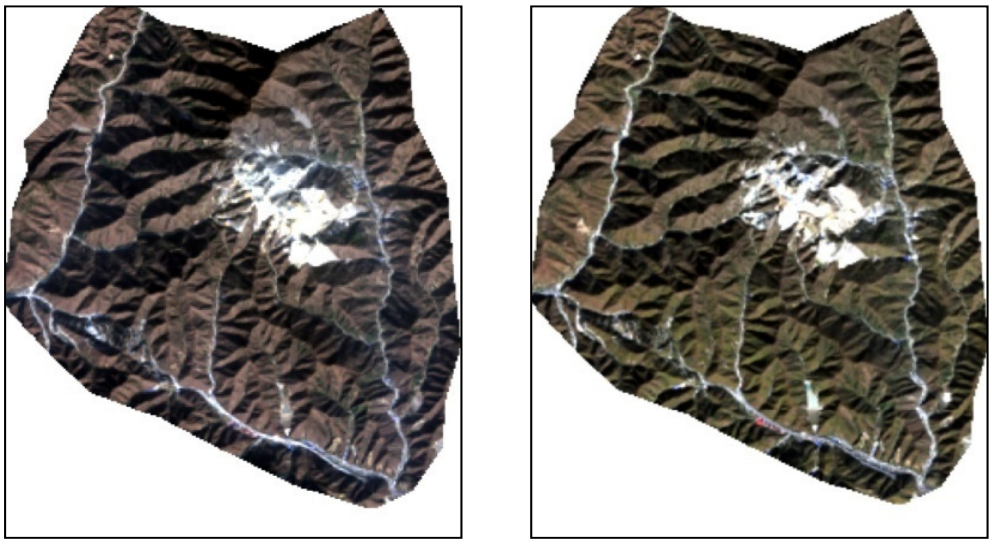
(**a**) (**b**)

Figure 2. Distribution of sampling points in the study area in (**a**) 2017 and (**b**) 2019.

The soil samples were processed and analysed in the laboratory of the environmental testing center of Guolian Quality Inspection Technology Co., Ltd. in China. The soil samples were crushed, dried, and passed through a nylon sieve. Other methods, such as flame atomic absorption spectrophotometer, were then used to characterise the concentrations of common heavy metal elements in all the processed samples. The soils were then studied for concentrations of Cu for subsequent use in the study, based on the extent to which the detected values of heavy metal elements in the samples exceeded the background values of heavy metal content in the soil, and the enrichment level of metals in the Daxigou-Yindongzi polymetallic deposit. A histogram analysis of Cu content derived from field-measured soil samples collected in 2017 and 2019, respectively, was performed to eliminate outliers.

Landsat 8 Multispectral Image and GIS Data

A radiometric correction on Landsat 8 multispectral image, followed by a geometric correction was performed. The corrected Landsat 8 OLI image is shown in Figure 3.



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

Figure 3. Landsat 8 OLI images (**a**) in 2017 and (**b**) in 2019 with radiometric correction.

In addition, the spatial distribution of heavy metals in soil is influenced by various factors, primarily terrain, human activity, and physical, chemical, and thermal factors. As stated in [7], using remote sensing images alongside other data as auxiliary information is necessary. However, acquiring additional auxiliary data is challenging. According to previous studies [11,23], auxiliary data, including DEM and some other GIS data, are beneficial in describing the spatial distribution of soil heavy metal concentrations. Therefore, in this study, DEM data and GIS auxiliary data related to human activities in the area were integrated with Landsat 8 satellite imagery.

2.2. Methodology

2.2.1. Analysing Factors Contributing to Cu Concentrations

Factors Derived from Landsat 8 Imagery

Vegetation coverage and clay minerals indirectly reflect the concentrations of heavy metals in the soil [11,23,24]. According to Ref. [24], many studies have shown that good correlations exist between vegetation indices and heavy metal concentrations. This is primarily due to the vegetation's root uptake of water and nutrients, which also include heavy metals. Certain plants, like hyperaccumulators, can accumulate high concentrations of heavy metals within their tissues, thereby reducing the overall soil concentration. Clay minerals can also effectively adsorb heavy metal ions in the soil, reducing their mobility [25].

According to the arguments, the paper proposes a combination of Landsat 8 spectral reflectance data and a variety of spectral indices and topographic data, alongside factors affecting human activities (e.g., distance to mining areas and distance to roads). The approach was intended to enhance the spatial characterisation of heavy metals in the soil, reducing issues largely caused by the lack of spectral information from Landsat 8 images. Additionally, the normalised difference vegetation index (NDVI), enhanced vegetation index (EVI), soil-adjusted vegetation index (SAVI), and clay mineral ratio (CMR) were selected as spectral indices. These spectral indices are listed in Table 1.

Table 1. Definition and application of vegetation indices based on Landsat 8 multispectral images.

Spectral Index	Formula	Description
NDVI	$(B5 - B4)/(B5 + B4)$	Describes the growth status and coverage of vegetation
EVI	$2.5 \times (B5 - B4)/(B5 + 6 \times B4 - 7.5 \times B2 + 1)$	More sensitive to vegetation cover than NDVI
SAVI	$1.5 \times (B5 - B4)/(B5 + B4 + 0.5)$	Reduces the impact of soil background compared to NDVI
CMR	B6/B7	Enhances rock mineral information in cohesive soil

In Table 1, B2—Blue band, B4—red band, B5—near infrared band, B6 and B7 represent two short wave infrared bands (SWIR1 and SWIR2) of Landsat 8 images. The four spectral indices derived from Landsat 8 image are shown in Figure 4.

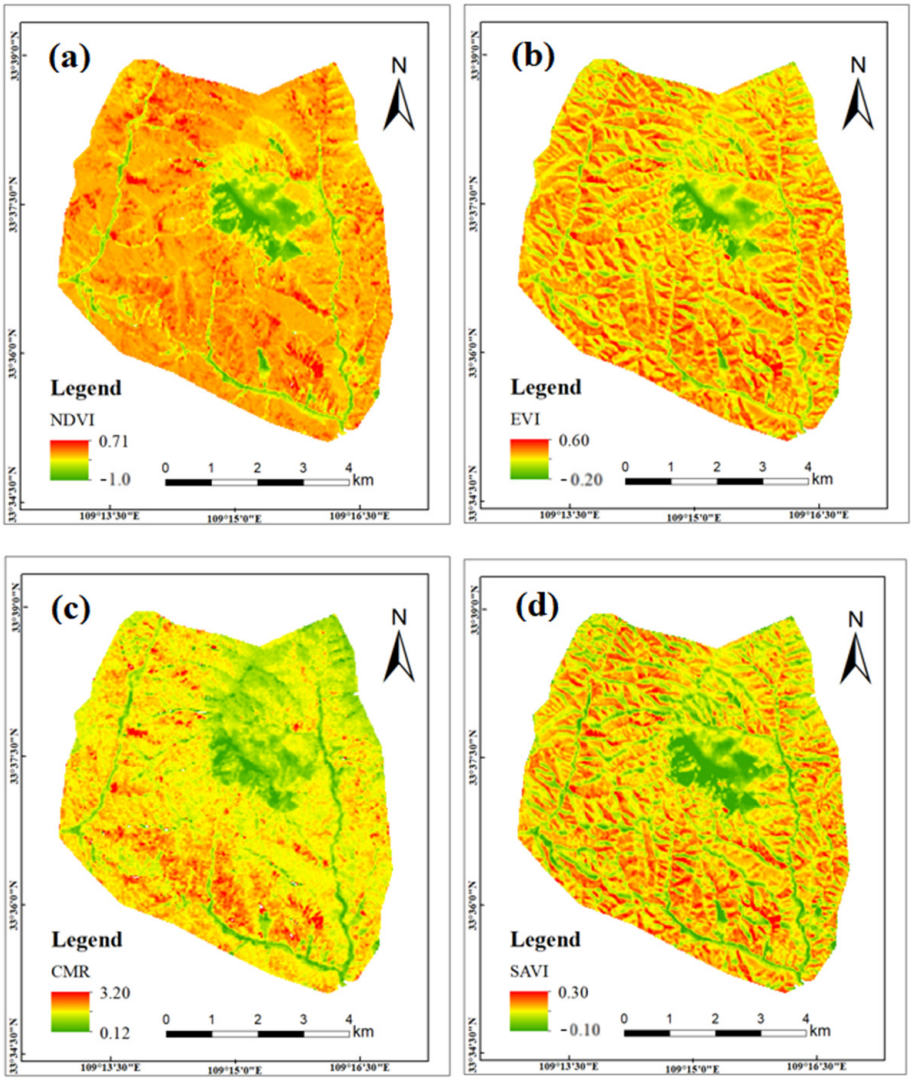


Figure 4. Spectral indices (**a**) NDVI; (**b**) EVI; (**c**) CMR; (**d**) SAVI from Landsat 8 images.

Figure 4a,b,d, show that most of the study area was densely vegetated, with the mining area and roads the least covered. Bare soil was observed in areas of sparse vegetation. The NDVI, EVI, and SAVI vegetation indices enhanced the resolution so that the bare soil areas could be characterised. The index of CMR enhances the rock mineral information in bare land and cohesive soil, which is beneficial for identifying vegetation cover areas, roads, and rock mineral enrichment areas. Figure 4c shows that there are significant differences in CMR values between the mining areas, ore stacking areas, mining roads, and forestland.

Factors Derived from GIS Auxiliary Data

According to previous studies, e.g., [7,11,23,24], some auxiliary data can be supplementary to spectral information of remote sensing imagery. In the paper, we introduced four factors, shown in Figure 5a–d, derived from DEM data and a land use map, favourable for retrieving heavy metal content in topsoil as shown in the following:

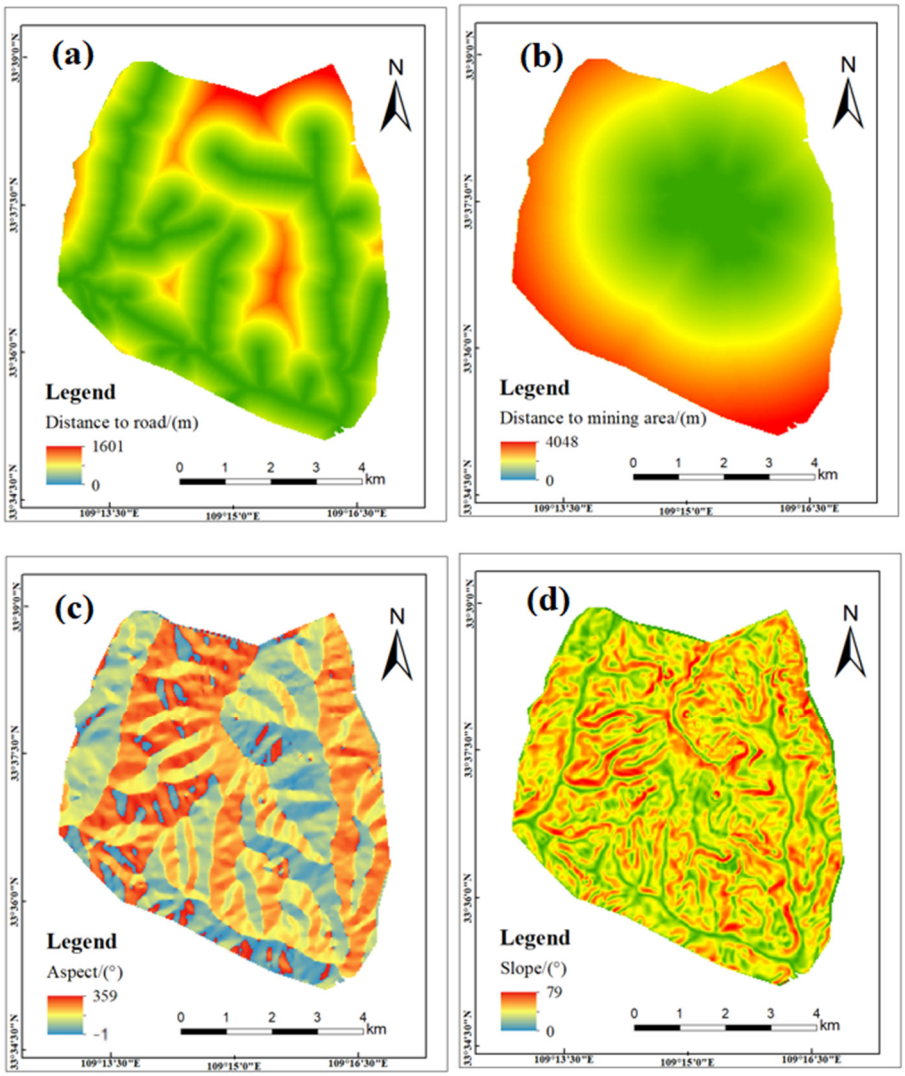


Figure 5. Human activities and terrain factors (**a**) Distance to road; (**b**) Distance to mining area; (**c**) Aspect; (**d**) Slope.

(1) Distance to road: the higher concentrations of heavy metals in the soil along the roads in the mining area because the ores and slag were dropped by vehicles carrying the ores. (2) Distance to mining area. The closer to the mining area, the higher the heavy metal content in the soil, and vice versa. (3) Aspect: it can reflect the differences in soil temperature. Temperature changes the form and distribution of heavy metals in the soil by affecting soil solid–liquid surface reactions, soil physicochemical properties, microbial processes, and so on. Reflecting differences driven by the orientation of each slope. (4) Slope: the slopes around the valley in the study area were relatively low, whereas other areas exhibited varying slopes, enhancing differences in terrain undulation and steepness in the area.

In order to analyse the correlation between those auxiliary data and Cu content in soil samples, in this study, we performed a correlation analysis on the five auxiliary variables, namely, the elevation, slope, aspect, distance to the mining area, and distance to the road in the mining area, with Cu content in topsoil, as shown in Table 2.

Table 2. Correlation analysis of auxiliary data with Cu content in topsoil.

Factor Type	Auxiliary Data	Coefficients
		Cu
Terrain-related	Elevation	0.490
	Slope	−0.041
	Aspect	0.300
Human activity related	Distance to mining area	−0.499
	Distance to road in the mining area	−0.103

Note: Here, the negative sign "−" in Table 2 represents a kind of negative correlation relationship that exists between some of the influence factors and the concentrations of the Cu in topsoil.

As seen in Table 2, the elevation, aspect, and distance to the mining area reached significant or extremely significant levels. Among them, the distance to the mining area reached a highly significant level, with the highest negative correlation coefficient of 0.499 for Cu. This indicates that the closer the site is to the mining area, the higher the content of the Cu element in the topsoil. According to the field survey, tramcars frequently traverse the roads in the mining area, resulting in the accumulation of mineral dust, which ultimately leads to an increase in the heavy metal content in the soil. Elevation values of the mining area also exhibit highly significant correlations with Cu concentrations, reaching a positive correlation value of 0.490. This indicates that the content of the Cu increases with a rise in the altitude, which is related to the terrain conditions of the mining area: samples with high heavy metal concentrations are typically found near the mining area, where the elevation is high. The elevation value indirectly reflects the relationship between the concentrations of heavy metals in soil and the distance to the mining site. The aspect can inhibit rain from washing away the heavy metals in soil towards the bottom of the mountain depending on the circumstance of vegetation growth. As shown in Table 2, the aspect is correlated with heavy metal content, however, the correlation coefficient is lower than that of elevation and distance to the mining area.

A visualisation of the relationship between the elevation, aspect, distance to the mining area (in short, Dis_MA), NDVI, and Cu concentrations in topsoil is useful for a more intuitive understanding of their relationships, as shown in Figure 6.

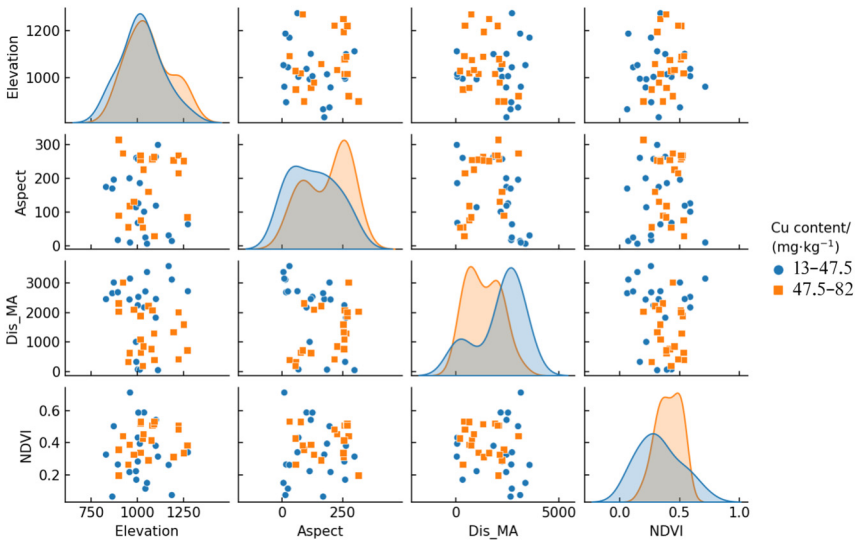


Figure 6. Scatter diagrams of auxiliary data and Cu concentrations in topsoil.

Based on Figure 6, only the Dis_MA factor has a significant impact on the spatial distribution difference of the Cu content; that is, the higher the content of Cu at a sampling site, the lower the Dis_MA value. In addition, the correlation between the above several factors is not apparent. This indicated that the conclusions drawn from the scatter plot

in Figure 6 align closely with those of the correlation analysis in Table 2, and that it can more intuitively reflect the correlation between Cu concentrations and various influencing factors. In conclusion, the terrain-related factors and human activity-related factors in the mining area in topsoil are favourable to retrieving Cu concentrations in the topsoil. This is in addition to the use of spectral indices, such as NDVI, as highlighted in this study.

2.2.2. Our proposed Model for Cu Retrieval in the Topsoil

The Traditional TrAdaboost Algorithm

AdaBoost is a boosting algorithm commonly used for different classification tasks [12]. The basic idea was to continuously update the sample weights during the learning process of the classifier parameters to generate different training sets. This assigns different weights based on the correct and incorrect classification of the samples. Specifically, the weights of the correctly classified samples were reduced, whereas those of incorrectly classified ones were increased. The results of each classifier were weighted to obtain the final results. However, when the samples are spatiotemporally distinct, the probability distribution of the data varies, resulting in the sample weight update mechanism of the AdaBoost algorithm exhibiting poor adaptability. Transfer learning involves transferring the information or knowledge learned in the source domain to the target domain. Among them, sample (instance) transfer learning selectively transfers samples from the source domain to the target domain. In the training process of target model, different weights are assigned to the samples from source or target domain, which can achieve an effective transfer of information, such as cross-domain samples.

TrAdaBoost is an AdaBoost algorithm based on sample transfer learning [26] that uses a different sample weight update method than AdaBoost. The TrAdaBoost regression algorithm updates the weights of samples based on the error between the predicted and measured values during the iteration process. In response to the issue of "even small errors can lead to a decrease in weight", in the sample update method of traditional TrAdaBoost regression algorithms, Pardo et al. [27] proposed a two-stage sample weight update the TrAdaBoost algorithm. The algorithm entails the gradual reduction of the source domain sample weight to a certain threshold during the training process. The weights of the source samples are frozen, and the weights of the target samples are updated.

Our Cross Time-Domain Transfer Learning Model Based on TrAdaboost for Cu Concentrations

We used historical sample information measured in 2017 and prior information on the soil heavy metal spatial distribution in the area, to compensate for the shortage of training sample data in 2019. Therefore, a cross-time-domain transfer learning model based on TrAdaboost algorithm for Cu concentration retrieval in the topsoil in 2019 was established based on the TrAdaBoost algorithm proposed in [27]. The model is described as follows:

The training dataset $D_{\mathrm{S}} = \{x_i, y_i\}_{i=1}^{n}$ in the source domain was acquired in 2017, where $x_i (i \in [1, \mathrm{n}])$ is the feature vector of the training samples and y_i is the corresponding measured value, with a probability distribution of $P(x_{\mathrm{S}})$. The dataset $D_{\mathrm{T}} = \{x_i, y_i\}_{i=\mathrm{n}+1}^{\mathrm{m}}$ in the target domain was acquired in 2019. It was divided into a training dataset $D_{\text{T-Train}}$ and a test dataset $D_{\text{T-Test}}$. $P(x_{\mathrm{T}})$, $P(x_{\mathrm{T}}) \neq P(x_{\mathrm{S}})$ is the probability distribution of the dataset.

Given a total training dataset $D_{\mathrm{STrain}} = D_{\mathrm{S}} \cup D_{\text{T-Train}}$ with a size of n + m, a decision tree was chosen as the base learner of AdaBoost model, we constructed a crosstime-domain sample transfer learning model based on the TrAdaBoost algorithm to predict the Cu concentrations of unknown samples in target domain.

We defined a weight $\mathrm{W}^t = \mathrm{W}_{\mathrm{S}}^{\mathrm{t}} \cup \mathrm{W}_{\mathrm{T}}^{\mathrm{t}}$ using the total training set $D_{\text{ST-Train}}$ at tth iteration for the training of our model, where $\mathrm{W}_{\mathrm{S}}^{\mathrm{t}} = \left(w_1^t, w_2^t, \ldots, w_n^t\right)$ is the weight vector of the training data in the source domain and $\mathrm{W}_{\mathrm{T}}^{\mathrm{t}} = \left(w_1^t, w_2^t, \ldots, w_{\mathrm{m}}^t\right)$ in the target domain. The parameter training algorithm of our model can be described as follows:

Input: The source dataset Ds and the target dataset D_{T}.

Set the following parameters, such as the total number of execution steps S, the maximum iteration of the learning process N, and the K-fold of cross-validation;

Normalise the source dataset *D*s and the target dataset D_{T};

Input the training dataset $D_{\text{ST-Train}}$ and initialise the weight of each training sample as $w_i^1 = 1/(\mathrm{n} + \mathrm{m})$, $\mathrm{i} \in [1, \mathrm{n} + \mathrm{m}]$;

For t = 1 to S do

Empty the base learner f_t and normalise the weights of each training sample;

Train the parameters of the model and update the weight $\mathrm{W}_{\mathrm{T}}^t$ using TrAdaBoostR2, while keeping the weight $\mathrm{W}_{\mathrm{S}}^t$ unchanged;

For k = 1 to K do

Construct a subset of $D_{\text{ST-Train}}^k$ of $D_{\text{ST-Train}}$;

Train the model using training dataset $D_{\text{ST-Train}}^k$ and calculate the predicted value of test data using the TrAdaBoost algorithm;

For i = 1 to m + n do

Calculate the relative error e_i^t of ith training sample using Equation (1),

$$e_i^t = \left| y_i - \left(f_i^t\right)^k \right| / e_{max}^t, \text{where } e_{max}^t = max_{i=\mathrm{n}+1}^{\mathrm{n}+\mathrm{m}} \left| y_i - \left(f_i^t\right)^k \right| \tag{1}$$

where $\left(f_i^t\right)^k$ is the predicted value of Cu concentration of the ith training sample using the tth base learner in the process of the kth fold cross-validation. y_i is the actual measured value of the ith sample;

Calculate the average error e^t over later n + m samples using Equation (2),

$$e^t = \sum_{i=n+1}^{n+m} \frac{w_i^t e_i^t}{\sum_{i=n+1}^{n+m} w_i^t} \tag{2}$$

Define the weight β_t of the base learner as Equation (3)

$$\beta_t = e^t / (1 - e^t), \tag{3}$$

Update the weight w_i^{t+1} of the ith sample using Equation (4) by minimising e^t based on the principle while keeping $\sum_{i=\mathrm{n}+1}^{\mathrm{m}} w_i^t = \frac{\mathrm{m}}{\mathrm{n}+\mathrm{m}} + \frac{t}{\mathrm{S}-1}\left(1 - \frac{\mathrm{m}}{\mathrm{n}+\mathrm{m}}\right)$ unchanged in the process of cross-validation;

$$w_i^{t+1} = \begin{cases} w_i^t \beta_t^{e_i^t} / Z_t, 1 \le i \le n \\ w_i^t / Z_t, n+1 \le i \le n+m \end{cases}, where\ Z_t = \sum_{i=1}^{n+m} w_i^t, \tag{4}$$

Resample $\mathrm{X} = \{x_i\}_{i=1}^{\mathrm{n}+\mathrm{m}}$ using Bootstrap many times;

Update the weight W^t of the training samples in $D_{\text{T-Train}}$ using the base learner;

Output: The optimal weight W^* of the training dataset, the optimal weight β^* of the base learner.

In this process, the decision tree was selected as the basic learner for regression. Weights of each learner and each sample were constantly adjusted and updated in the process of training using the K-fold cross validation (CV) method. With an optimal weights W^* and the optimal weight β^* of weak learners, a final strong learner f_T can be calculated using weighted average method, and our sample transfer learning model was constructed. Then, the value of Cu concentration at each unknown site in the study area can be predicted using Equation (5),

$$\mathrm{y}_j^{pred} = \mathrm{model}(\mathrm{f}_T, \mathrm{W}^*), \tag{5}$$

The main parameters were set as follows: a decision tree (the maximum depth was 4) was selected as the base learner. The maximum number of iterations is 50. The learning rate was set to 0.01. The loss function is linear. The initial weight was set as the reciprocal

of the total number of training sets, i.e., 1/46 here. The total iterations S was 10. The cross-validation fold is 5. For our cross-time-domain transfer learning model mentioned above, training samples selected from samples collected in 2017 and 2019 with different probability distribution.

2.2.3. Evaluation Metrics

Three measures, including the root-mean-square deviation (RMSE), goodness of fit (R^2), and relative prediction deviation (RPD) [24], were used to evaluate the performance of our model. RMSE measures the deviation between the observed and measured values, and the smaller its value, the higher the inversion accuracy of the model. RPD reflects the predictive ability of the model, and R^2 reflects the fitting effect of the model, with a value range of [0, 1]. A higher R^2 is indicative of a better-fitting model. The calculation formulae for the abovementioned indicators are shown in Equations (6)–(8).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{N}}, \tag{6}$$

$$RPD = \frac{\sqrt{\sum_{i=1}^{N} (y_i - \overline{\hat{y}_i})^2}}{(N-1)^2 RMSE}, \tag{7}$$

$$R^2 = \frac{\sum_{i=1}^{N} (y_i - \overline{\hat{y}_i})^2}{\sum_{i=1}^{N} (\hat{y}_i - \overline{\hat{y}_i})^2}, \tag{8}$$

In Equations (6)–(8), y_i is the measured value of heavy metal concentrations (regarded as the true value), $\hat{y}_i$ is the predicted value of heavy metal concentrations, $\overline{\hat{y}_i}$ is the average of all predicted values, and N is the total number of samples.

3. Results and Validation

3.1. Experiments

To evaluate the performance of the model proposed in this study, three groups of the experiment were designed based on different training sample conditions, which were compared with the traditional Adaboost [12] and the TrAdaBoost algorithm [27] under different training strategies, as shown in Table 3.

Table 3. Experimental settings.

Group	Item	Purpose
A	Training samples from only in 2019	To evaluate the performance of traditional Adaboost model in case of the same probability distribution but fewer training samples.
B	Training samples from all samples collected in 2017 and 2019	To evaluate the performance of traditional TrAdaboost model in case of different probability distribution but all of training samples.
C	Training samples selected from samples collected in 2017 and 2019	To evaluate the performance of our TrAdaboost model in case of different probability distribution and a selective use of training samples by transfer learning.

In the experiment group A, B, and C, the main parameters were same (details can be seen in Section Our Cross Time-Domain Transfer Learning Model Based on TrAdaboost for Cu Concentrations). Our algorithm mentioned above was implemented in Windows 10 using Python 3.7 programming language, and the PyCharm environment.

3.2. Results and Validation

The accuracy of retrieving Cu concentrations in 2019 was evaluated using test set collected in the field. To furthermore analyse the performance of our proposed model in group C, the contrast to others in groups A and B were shown in Table 4.

Table 4. Comparison of the accuracy of our model to the two other experiments.

Group	Number of Samples		Evaluation of Training Data			Evaluation of Test Data		
	Training	Test	R^2	RMSE	RPD	R^2	RMSE	RPD
A	31	11	0.82	7.30	1.95	0.60	7.84	1.64
B	46	11	0.74	11.78	1.93	0.55	8.33	1.37
C	46	11	0.78	10.83	2.51	0.66	7.24	1.76

As shown in Table 4, we can see 31 training samples from the same period in 2019 were used to train the model in group A, and the R^2 value of the training set reached 0.82 due to the same distribution as that of the test set, while the R^2 value of the test set was 0.60. This indicates that the traditional AdaBoost regression model exhibits over-fitting under these conditions; In group B, the training set was directly expanded, and 15 training data points from 2017 were randomly added to the 2019 training data. We found that the R^2 value for the test set decreased from 0.60 to 0.55, and the accuracy of the training set also decreased. It showed that traditional AdaBoost regression models do not perform well (RPD $\leq$ 1.5); In group C, our model was used to selectively transfer the samples from 2017. Consequently, the R^2 value of the test set increased to 0.66, the RPD was 1.76, and the RMSE was 7.24 mg·kg^{-1}. This indicates that the model reached an approximate standard.

Therefore, the following conclusions can be drawn: the accuracy of our proposed model was not impacted by the samples driving the negative transfer in the source domain. The selective use of effective training samples in the source domain improved the regression accuracy of the model. It showed that the performance of our model on the test set data has been improved compared with that of group A and B.

The mapping of the Cu concentrations in the study area using our proposed model as in group C in the Tables 3 and 4 was shown in Figure 7.

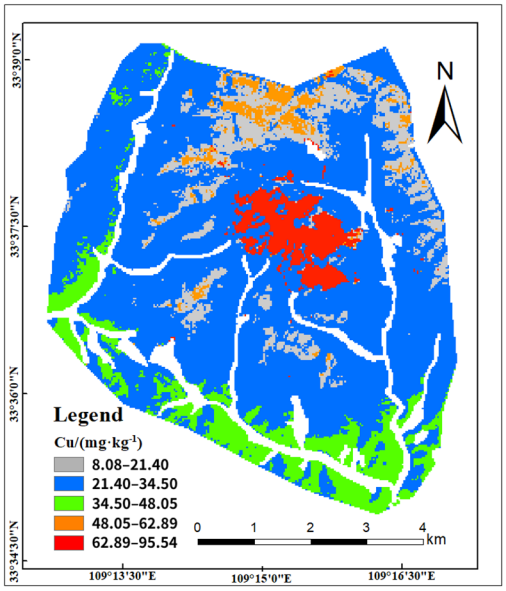


Figure 7. The spatial distribution of Cu concentrations using our proposed model.

Figure 7 shows that the area was filled with higher concentrations of Cu, mainly in the mining area and its surrounding areas. Cu was distributed in the soil along the roads in

the northern and southern parts of the mining area, as the ores and slag were dropped by vehicles carrying the ores. According to Ref. [28], the screening value of soil contamination risk on agricultural land for Cu content is 50 mg·kg^{-1}. From Figure 7, Cu concentrations in the mining area (red colour) and a part of its out-ring (ore stacking area) exceeded the risk value of soil contamination on agricultural land, but did not exceed the state standard of soil environment background value in China. Thus, soil environment monitoring in the area should be strengthened. In addition, this spatial distribution pattern is consistent with the spatial distribution patterns of heavy metals obtained in previous studies [4,11]. Note that a masking treatment was applied (white area) owing to the lack of soil on roads and buildings in the study area.

4. Discussion

From Table 4, the R^2 value for the training set is mainly better than that of the test set, which shows the generalisation of our model is not enough. One possible reason is that the data distribution between the training and the test sets is inconsistent, resulting in insufficient generalisation ability of the model on the test set. In addition, it is relevant to an inappropriate parameter setting, an insufficient dataset, and sensitivity to noise from the dataset. However, the results showed that the accuracy of the regression model on the test set data improved compared to two other groups.

To furthermore analyse and compare the performance of our proposed model in group C with A, B, the estimated values of the model in this study were fitted to the measured values, as shown in Figure 8a–c.

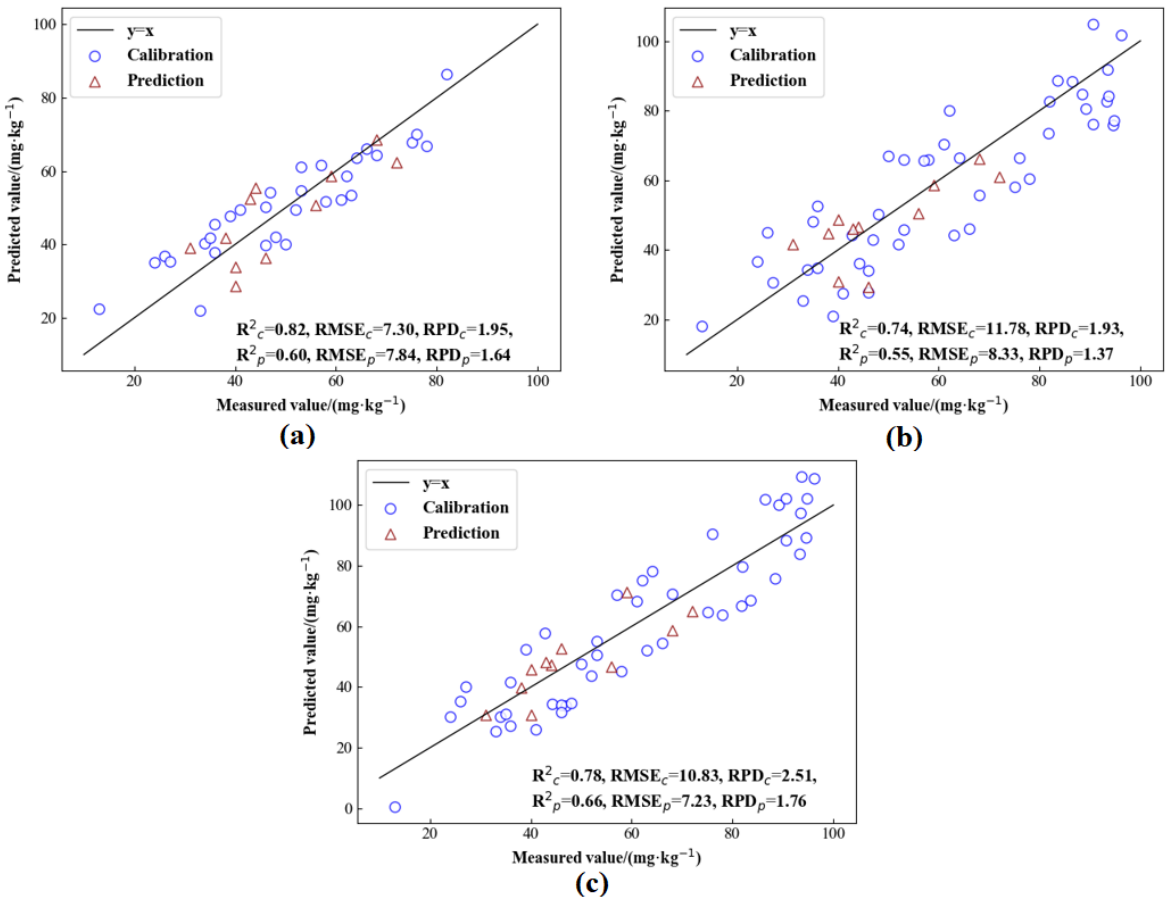


Figure 8. Prediction results of our model under the group A (**a**), group B (**b**), and group C (**c**).

Based on a comprehensive analysis of Table 4 and Figure 8, the following conclusions can be drawn:

In group A, the test set scatter points along the y = x line exhibited a relatively larger deviation than those of the training set in the scatter plot. According to the evaluation criteria of the model's inversion ability [24], this indicates that the model has ordinary ability in regression of heavy metal content ($0.5 \leq R2 \leq 0.65$ or $1.5 \leq RPD \leq 2.0$) but does not meet the approximate standard of the regression model ($0.65 \leq R2 \leq 0.81$ or $2.0 \leq RPD \leq 2.5$).

In group B, the structure of the scatter plot represents the negative impact associated with the cross-time-domain transferring of samples (i.e., training samples). This is because of the significant difference in the probability distribution between the training and test datasets. Therefore, either indiscriminately combining training sample data with different probability distributions or setting the same weight for each sample is not conducive to improving model accuracy.

In group C, the training data were the same as those of group B, however, the results showed that the accuracy of our model on the test set data was improved compared with the results of groups A and B. It proved that the selective use of effective training samples in the source domain improved the regression accuracy of the model.

5. Conclusions

We propose a cross-time-domain transfer learning model to monitor heavy metal pollution based on the traditional TrAdaBoost algorithm under the condition of a few samples, which realised the effective transfer and reuse of samples collected in the historical period in 2017 to the new monitoring period in 2019. This approach avoided problems associated with sample collection costs and improved the efficiency at which long-term soil monitoring efforts (i.e., within the context of heavy metals) can be conducted. By comparing and analysing the regression accuracy of the model with the measured sample data, several important findings were derived. The proposed cross-time-domain transfer learning model, based on the condition of few samples, can selectively use sample data obtained in the past to improve the regression accuracy of soil heavy metal content in the new monitoring period when the number of training samples is limited. Additionally, directly combining source domain samples with target domain samples does not improve the model prediction accuracy. We found that the proposed model accurately estimated the content of Cu in the topsoil of the mining area in 2019 and was more accurate than the traditional TrAdBoost algorithm. The model holds potential for more accurate and cost-effective monitoring of heavy metal pollution in soil, facilitating remediation-related efforts. The spatial distribution of Cu concentrations in the study area obtained from this model was consistent with those derived from previous studies, which further validates the effectiveness of the proposed model in the study.

Author Contributions: Conceptualization, Y.Y. (Yun Yang); Methodology, Y.Y. (Yun Yang), Q.T. and H.B.; Software, Q.T.; Validation, H.B.; Formal analysis, Y.W. and Y.Y. (Yi Yan); Investigation, A.H.; Resources, Y.W. and A.H.; Data curation, Q.T. and Y.Y. (Yi Yan); Writing—original draft, H.B.; Writing—review and editing, Y.Y. (Yun Yang), Q.T., Y.W., Y.Y. (Yi Yan) and A.H.; Visualization, Q.T.; Supervision, Y.Y. (Yun Yang); Project administration, Y.Y. (Yun Yang); Funding acquisition, Y.Y. (Yun Yang). All authors have read and agreed to the published version of the manuscript.

Funding: This research was jointly funded by Natural Science Basic Research Program of Shaanxi in China (No. 2022JM-163), the Basic Scientific Research Business of Central University of Chang'an University (No. 300102269205) and the National Natural Science Foundation of China (No. 41301386, No. 42261144749, No. 42377158).

Data Availability Statement: Our observed data in field is unavailable because of privacy.

Acknowledgments: The authors are grateful to NASA Land Processes Distributed Active Archive Center User Services, USGS Earth Resources Observation and Science (EROS) Center for providing the Landsat 8 data, and the geospatial data cloud in China for providing DEM data.

Conflicts of Interest: The Author Han Bai was employed by the company State Grid Location Based Service. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Namee, A.M.; Bahaa, Z.; Fattah, M.Y. Some Strategies for Reducing and/or Removing Heavy Metals from Contaminated Soil: A Review. *Proc. AIP Conf.* **2023**, *2775*, 030006. [CrossRef]
2. Vuong, X.T.; Vu, L.D.; Duong, A.T.T.; Duong, H.T.; Hoang, T.H.T.; Luu, M.N.T.; Nguyen, V.D.; Nguyen, T.T.T.; Van, T.H.; Minh, T.B. Speciation and environmental risk assessment of heavy metals in soil from a lead/zinc mining site in Vietnam. *Int. J. Environ. Sci. Technol.* **2022**, *20*, 5295–5310. [CrossRef]
3. The Ministry of Environmental Protection. *The Ministry of Land and Resources*; National Soil Contamination Survey Report; The Ministry of Environmental Protection: Beijing, China, 2014.
4. Bai, H.; Yang, Y.; Cui, Q.F.; Jia, P.; Wang, L. Retrieval of heavy metal concentrations in soil using GF-5 satellite images based on GA-XGBoost model. *Laser Optoelectron. Prog.* **2022**, *59*, 525–534. [CrossRef]
5. Zhang, B.; Guo, B.; Zou, B.; Wei, W.; Lei, Y.; Li, T. Retrieving soil heavy metals concentrations based on GaoFen-5 hyperspectral satellite image at an opencast coal mine, Inner Mongolia, China. *Environ. Pollut.* **2022**, *300*, 118981. [CrossRef] [PubMed]
6. Cheng, Y.S.; Zhou, Y. Research progress and trend of quantitative monitoring of hyperspectral remote sensing for heavy metals in soil. *Trans. Nonferrous Met. Soc. China* **2021**, *31*, 3450–3467. [CrossRef]
7. Bou Kheir, R.; Shomar, B.; Greve, M.B. 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]
8. Ghrefat, H.; Zaman, H.; Batayneh, A.; El Waheidi, M.M.; Qaysi, S.; Al-Taani, A.; Jallouli, C.; Badhris, O. Assessment of heavy metal contamination in the soils of the gulf of aqaba(north western saudiarabia): Integration of geochemical, remote sensing, gis, and statistical data. *J. Coast. Res. Int. Forum Littoral Sci.* **2021**, *4*, 37. [CrossRef]
9. Yu, J.; Wang, J.; Leblon, B.; Song, Y. Nitrogen Estimation for Wheat Using UAV-Based and Satellite Multispectral Imagery, Topographic Metrics, Leaf Area Index, Plant Height, Soil Moisture, and Machine Learning Methods. *Nitrogen* **2021**, *3*, 1–25. [CrossRef]
10. Liu, M.; Liu, X.; Wu, M.; Li, L.; Xiu, L. Integrating spectral indices with environmental parameters for estimating heavy metal concentrations in rice using a dynamic fuzzy neural-network model. *Comput. Geosci.* **2011**, *37*, 1642–1652. [CrossRef]
11. Wang, T.J.; Zhao, M.H.; Yang, Y.; Cui, Q.; Li, L. Inversion of heavy metals concentrations in soil using multispectral remote sensing imagery in Daxigou mining area of Shaanxi. *Spectrosc. Spectr. Anal.* **2019**, *39*, 3880–3887.
12. Cao, Y.; Miao, Q.G.; Liu, J.C.; Gao, L. Advance and prospects of AdaBoost algorithm. *Acta Autom. Sin.* **2013**, *39*. [CrossRef]
13. Yuan, Z.R.; Wei, L.F.; Zhang, Y.X.; Yu, M.; Yan, R.R. Hyperspectral Inversion and Analysis of Heavy Metal Arsenic Content in Farm land Soil Based on Optimizing CARS Combined with PSO-SVM Algorithm. *Spectrosc. Spectr. Anal.* **2020**, *40*, 567–573. [CrossRef]
14. Zhang, Z.H.; Guo, F.; Xu, Z.; Yang, X.Y.; Wu, K.Z. On retrieving the chromium and zinc concentrations in the arable soil by the hyperspectral reflectance based on the deep forest. *Ecol. Indic.* **2022**, *144*, 109440. [CrossRef]
15. Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A Comprehensive Survey on Transfer Learning. *Proc. IEEE* **2021**, *109*, 43–76. [CrossRef]
16. Zhang, Y.; Guo, X.; Leung, H.; Li, L. Cross-task and cross-domain sar target recognition: A meta-transfer learning approach. *Pattern Recognit. J. Pattern Recognit. Soc.* **2023**, *113*, 109402–109412.
17. Tao, C.; Cui, W.B.; Wang, Y.J.; Zou, B.; Zou, Z. Soil heavy metal qualitative classification model based on hyperspectral measurements and transfer learning. *Spectrosc. Spectr. Anal.* **2019**, *39*, 2602–2607. [CrossRef]
18. Liu, L.F.; Ji, M.; Buchroithner, M. Transfer learning for soil spectroscopy based on convolutional neural networks and its application in soil clay concentrations mapping using hyperspectral imagery. *Sensors* **2018**, *18*, 3169. [CrossRef] [PubMed]
19. Tao, C.; Wang, Y.J.; Cui, W.B.; Zou, B.; Zou, Z.; Tu, Y. A transferable spectroscopic diagnosis model for predicting arsenic contamination in soil. *Sci. Total Environ.* **2019**, *669*, 964–972. [CrossRef] [PubMed]
20. Wang, Y.; Tao, C.; Zou, B. A transfer learning approach utilizing combined artificial samples for improved robustness of model to estimate heavy metal contamination in soil. *IEEE Access* **2020**, *8*, 176960–176972. [CrossRef]
21. Fang, W.X.; Hu, R.Z. Mineralization zoning in yindongzi-daxigou. *Chin. J. Geochem.* **2001**, *20*, 45–51. [CrossRef]
22. Chen, L.; Wang, J.D.; Gu, T.F.; Yu, G.Q.; Wang, A.G. Research on fuzzy comprehensive evaluation of ecological environment in Daxigou iron mine. *Chin. J. Soil Sci.* **2017**, *48*, 794–799. [CrossRef]
23. Mei, X.; Liu, H.Y.; Wu, L.H.; Luo, Y.X.; Zhang, Q.Y.; Jing, P.; Faustino, D.; Song, W. Spatial distribution and influencing factors of heavy metals in soil of northwest Guizhou based on HDXRF and ICP-MS. *Soil* **2023**, *55*, 399–408. [CrossRef]
24. Shi, T.Z.; Chen, Y.Y.; Liu, Y.L.; Wu, G. Visible and near-infrared reflectance spectroscopy-An alternative for monitoring soil contamination by heavy metals. *J. Hazard. Mater.* **2014**, *265*, 166–176. [CrossRef] [PubMed]
25. Zhu, W.; Liu, D.H.; Chen, J.Q.; Dai, Q.Y.; Fu, Y.C. Research Progress on the Application of Clay Minerals in the Remediation of Cadmium Polluted Farmland. *Soil Bull.* **2018**, *49*, 499–504. [CrossRef]
26. Dai, W.Y.; Yang, Q.; Xue, G.R.; Yu, Y. Boosting for transfer learning. In Proceedings of the 24th International Conference on Machine Learning, Corvallis, OR, USA, 20–24 June 2007.

27. Pardoe, D.; Stone, P. Boosting for regression transfer. In Proceedings of the 27th International Conference on Machine Learning, DBLP, Haifa, Israel, 21–24 June 2010; Available online: https://icml.cc/Conferences/2010/papers/330.pdf (accessed on 10 May 2023).
28. *GB 15618-2018*; Soil environment Quality Risk Control Standard for Soil Contamination of Agriculture Land (Trial). China Environmental Publishing Group, Ministry of Ecological Environment: Beijing, China, 2018.