



Rapid detection and risk assessment of soil contamination at lead smelting site based on machine learning

Sheng-guo XUE¹, Jing-pei FENG¹, Wen-shun KE¹, Mu LI¹, Kun-yan QIU², Chu-xuan LI¹, Chuan WU¹, Lin GUO³

1. School of Metallurgy and Environment, Central South University, Changsha 410083, China;

2. Henan Key Laboratory of Monitoring and Remediation in Heavy Metal Polluted Soil, Jiyuan 454650, China;

3. Henan Academy of Geology, Zhengzhou 450000, China

Received 7 March 2023; accepted 5 September 2023

Abstract: A general prediction model for seven heavy metals was established using the heavy metal contents of 207 soil samples measured by a portable X-ray fluorescence spectrometer (XRF) and six environmental factors as model correction coefficients. The eXtreme Gradient Boosting (XGBoost) model was used to fit the relationship between the content of heavy metals and environment characteristics to evaluate the soil ecological risk of the smelting site. The results demonstrated that the generalized prediction model developed for Pb, Cd, and As was highly accurate with fitted coefficients (R^2) values of 0.911, 0.950, and 0.835, respectively. Topsoil presented the highest ecological risk, and there existed high potential ecological risk at some positions with different depths due to high mobility of Cd. Generally, the application of machine learning significantly increased the accuracy of pXRF measurements, and identified key environmental factors. The adapted potential ecological risk assessment emphasized the need to focus on Pb, Cd, and As in future site remediation efforts.

Key words: smelting site; potentially toxic elements; X-ray fluorescence; potential ecological risk; machine learning

1 Introduction

Potentially toxic elements (PTEs) in soil have attracted worldwide concern due to their generally high toxicity and non-degraded characteristics [1,2]. Mining and smelting activities are the main anthropogenic sources of PTEs in soil [3,4]. China, as the largest producer and consumer of non-ferrous metals globally, has amassed a significant number of PTEs in the soil surrounding non-ferrous metal smelting sites [5]. Consequently, accurate identification, assessment, and detection of multi-metals in contaminated sites are crucial for effective site remediation and contamination control.

Traditionally, the accurate PTEs content in the

soil has been determined by atomic absorption spectroscopy (AAS) or inductively coupled plasma optical emission spectrometry (ICP-OES) after concentrated acid digestion (see USEPA 3050b, 3051a, and 3052). Although these methods can accurately measure total PTEs content in the soil, the measurement process is time-consuming, laborious, and expensive. Portable X-ray fluorescence (pXRF) is a simple, fast, accurate measurement mean of determining element content. It offers several advantages over traditional methods, including low detection limits, multi-element detection capability, inexpensive, and environment friendly characteristics [6]. As a result, pXRF has been widely used as a rapid screening tool for contaminants in fieldwork at contaminated

Corresponding author: Wen-shun KE, E-mail: kewenshun@csu.edu.cn;

Lin GUO, E-mail: guolin_cug@126.com

DOI: 10.1016/S1003-6326(24)66595-7

1003-6326/© 2024 The Nonferrous Metals Society of China. Published by Elsevier Ltd & Science Press

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

sites. However, previous studies [7,8] have identified several factors leading to deviations between the measured results and the actual PTEs contents in the soil, including water content, organic matter content, soil particle size, soil matrices, and mineral morphology. Therefore, as an effective and convenient method to measure PTEs contents, it is necessary to improve the accuracy of pXRF measurement with a simpler method and make it suitable for different measurement environments and enormous sites.

PTEs in the soil can threaten human health through physiological/molecular pathways, and long-term exposure to PTEs contamination causes pathological changes in the human body [9,10]. Therefore, an accurate ecological risk assessment method is urgently needed. The potential ecological risk index (RI) has been widely used for the evaluation of soil PTEs pollution levels, which integrates the concentrations and toxicity of each PTE to provide a comprehensive assessment of the level of contamination. However, the toxicity and mobility of PTEs in soil mainly rely on their presence pattern [11,12], and the risk value assessment method relying solely on the total PTEs content may exaggerate the risk level of PTEs in soils. Therefore, an accurate ecological risk assessment method is needed to guide site remediation.

Machine learning (ML), a subset of artificial intelligence, can adapt and learn from multi-dimensional, complex, and large data to build predictive models and mine the inner relationship between data. ML has been successfully used to correct the heavy metal values measured by pXRF [13,14]. However, potential variations in soil properties among sites will affect PTEs forms. In addition to diversities in organic matter abundance and soil hydration status, there are also variances in heavy metal distributions [15]. Therefore, it follows that distinct pXRF measurement errors can arise at every location, underscoring the importance of understanding local contexts and potential sources of variation. Some substances in the soil may change their fugitive form in the soil by binding to PTEs. For example, iron oxides, an active soil constituent, exhibit a strong ability to adsorb PTEs [16], and phosphates tend to induce heavy metals to form phosphate minerals, especially Pb [17]. However, few reports discuss the pXRF

measurement biases caused by environmental features.

In this study, the objects were to predict actual heavy metal levels at a smelting site by utilizing ML algorithms that incorporate environmental characteristics (such as available phosphorus, crystalline iron, amorphous iron oxide, free iron oxide, and pH) and pXRF measured PTEs levels. A general model for seven PTEs was established by using the XGBoost algorithm in an innovative way to uncover the intrinsic link between PTEs and complex soil environmental factors. Based on this, a potential ecological risk evaluation method was established that incorporates relevant environmental conditions, providing a more comprehensive framework for site assessment.

2 Experimental

2.1 Study area and sampling

The study was conducted at an abandoned Pb smelter located in Central China with an area of about 15000 m² (Fig. 1). 32 sampling points were arranged by the grid distribution method, and JDL150 crawler probes were used to drill the subsurface soil cores. The upper 3 m layer was sampled every 0.5 m, the lower 3–6 m layers were sampled every 1 m, and 207 soil samples at different depths were collected. All soil samples were taken to the laboratory and naturally air-dried for 7 d.

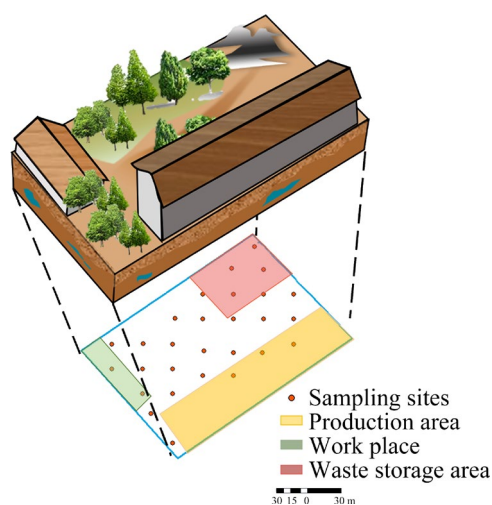


Fig. 1 Study area with locations of sample sites

2.2 Characterization of soil samples

To better discuss the effect of physicochemical properties of the soil on the measurement

process [18,19], the soil was air-dried and sieved (150 μm mesh) prior to testing with portable X-ray fluorescence (pXRF, Delta Premium XPD 600, Olympus Innov-X, USA). Standard samples were used to calibrate before measuring. Each sample was scanned for 60 s and the mean of the three measurements was calculated as the final result. The exact concentration of heavy metals in the samples was determined by the inductively coupled plasma optical emission spectrometer (ICP-OES, ICAP 7000 Series, Thermo Scientific, USA) after digestion extraction with a mixture of HNO_3 and HCl (1:3, volume ratio). To ensure the accuracy of the experimental analysis results, the method of two blanks and standard reference materials were used for each test batch (GBW07424–GBW07430). The recoveries of Cu, Ni, Pb, Cd, Zn, Sb, and As for the standard reference material were between 80% and 125%. The relative deviation between the determinate result and the actual concentration value was less than 10%.

Soil pH was measured using a pH meter in the extraction solution (ratio of solid/water is 1:5 (mg/mL)). The content of soil organic matter (SOM) was determined by the low-temperature external thermal potassium dichromate oxidation-colorimetric method. Soil available phosphorus (AP) was extracted with 0.5 mol/L sodium bicarbonate and analyzed by the molybdenum–antimony anti-colorimetric method. The free iron (Fe_d) oxide was extracted by the dithionite–citrate–bicarbonate (DCB) method; the amorphous iron (Fe_o) oxide was extracted by ammonium oxalate buffer. After extraction, the iron content was determined by the plasma emission spectrometer (VISTA-MPX). The crystalline iron (Fe_c) oxide content was calculated as a difference between free iron oxide content and amorphous iron oxide content.

2.3 Selection and optimization of ML model

Prior to the modeling, all pXRF results from scans of soil samples were normalized to improve the efficiency of the model, except this, there was no transformation in the data. Further, the total data were randomly divided into training set and validation set according to the proportion of 75% and 25%, respectively. Separate ML models were developed for two conditions: one using only

pXRF measurements and the other using pXRF measurements and environment characteristic values.

In this study, four traditional learning algorithms, including Linear Regression (LR), Random Sample Consensus (RANSAC), Decision Tree (DT) and Support Vector Machine (SVM), and two ensemble algorithms, including random forest (RF) and eXtreme gradient boosting (XGBoost), were used to train and predict the PTE contents. The introduction and parameter settings for each model were presented in the Supplement Materials.

2.4 Adapted potential ecological risk assessment

XGBoost was applied to analyzing the relationship between each environmental factor and PTEs, and determining the respective weights of each factor [20]. The environmental characteristic value corresponding to each PTE was calculated by the weight of each environmental characteristic and assigned to the potential ecological risk index after normalizing the environmental characteristic value, and the calculation formulas were as follows:

$$F_i = \sum_{i=1}^N W_i V_i \quad (1)$$

$$w_i = 1 - \frac{F_i - F_{\max}}{F_{\max} - F_{\min}} \quad (2)$$

$$\text{AE}_j^i = w_i E_j^i \quad (3)$$

$$\text{ARI}_j = \sum_{i=1}^N \text{AE}_j^i \quad (4)$$

where W_i denotes the weights of different environmental characteristics, N denotes the number of environmental characteristics, V_i denotes the normalized environmental characteristic value, F_i denotes the environmental characteristic value corresponding to each PTE, and the obtained environmental characteristic value is normalized to obtain the weight w_i reflecting the influence of environmental factors, AE_j^i is the adapted potential ecological risk value brought by PTE i at sampling site j , E_j^i denotes the single potential ecological risk index of PTE i at sampling site j , and ARI_j represents the adapted potential ecological risk index brought by the PTEs investigated in the study at sampling site j .

2.5 Statistical analysis

Linear Regression, Random Sample Consensus,

Decision Trees, Support Vector Machines, and Random Forests were all implemented using Python's Scikit-Learn machine learning package (version 1.0.2), and eXtreme Gradient Boosting was implemented using the XGBoost (1.6.1) package. Grid search was employed to optimize parameters for superior predictions. For each algorithm, only essential hyperparameters known to significantly impact the prediction process (see Supplementary Materials) were selected for parameter fine-tuning; other parameters remained fixed at their default settings.

To further ensure and improve the robustness and predictability of the model, 10-fold cross-validation was used to adjust the model parameters and evaluate whether there was overfitting. Two indices including the coefficient of determination (R^2) and root mean square error (RMSE) were used to evaluate the predictive performance of the model [21]. RMSE is commonly used to measure the deviation of the true value from the predicted value (Eq. (5)). The smaller the value of RMSE, the higher the accuracy of the model prediction.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - m_i)^2} \quad (5)$$

where n is the number of samples, y_i is the predicted value of the model, and m_i is the observed value of the chemical analysis.

SHAP interpretation is a local interpretation

method extended from the concept of Shapley value in game theory, which aims to distribute the contributions of players fairly in the game [22]. In ML, the Shapley value can quantify the contribution of each feature in the model. The SHAP (0.41.0) package in Python was used for visualization.

3 Results and discussion

3.1 Characteristics of contaminated smelting site soil

The contents of Ni, As, Cu, Pb, Zn, Cd, and Sb were determined by pXRF and ICP, as shown in Table 1. The average contents of Ni, Cu, Pb, Cd, and Sb measured by ICP were significantly higher than the average contents measured by pXRF, which was consistent with the results from XIA et al [23]. There was a metal-dependent discrepancy between pXRF and ICP measurements [16]. QU et al [24] found a similar conclusion that the mean and coefficient of variation (CV) of Cu measured by ICP were higher than those by pXRF, which may be due to the presence of organic matter and other substances in the soil. Notably, the pXRF values for Cu, Cd, and Sb exhibited similar median, mean, and standard deviations (SD) compared to those determined by ICP. The CV of PTEs measured by pXRF was almost not different from that measured by ICP, except Ni. The CV values of As, Cu, Pb, Zn, Cd, and Sb were all greater than 1, indicating that

Table 1 PTE contents measured by pXRF and ICP (mg/kg, $n=207$)

Item	ICP-Ni	ICP-As	ICP-Cu	ICP-Pb	ICP-Zn	ICP-Cd	ICP-Sb
Max	466	2130	1574	7790	1580	250	517
Min	8.47	1.61	25.7	16.6	3.30	0.014	0.18
Median	32.0	14.9	62.1	184	92.7	1.39	2.99
Mean	34.7	65.8	127.5	698	134	11.5	18.9
SD	36.1	183	193	1221	155	32.4	52.6
CV	1.04	2.78	1.51	1.75	1.16	2.81	2.78
Item	XRF-Ni	XRF-As	XRF-Cu	XRF-Pb	XRF-Zn	XRF-Cd	XRF-Sb
Max	154	2664	1220	7314	1962	204	133
Min	2.71	0.01	7.37	8.47	27.3	0.02	0.11
Median	33.5	18.7	63.5	108	101	1.11	3.05
Mean	34.1	102	132	526	161	9.91	20.7
SD	16.1	303	184	1020	227	25.9	58.3
CV	0.47	2.97	1.39	1.94	1.42	2.62	2.82

these PTEs had great chances of being affected by external causes such as anthropogenic activities. The spatial distribution of PTE contents in the soil exhibited strong variability.

The Pearson correlation heat map between pXRF and ICP measurements of PTEs was shown in Fig. 2. An exceptionally strong correlation was observed between pXRF and ICP contents of Cd, Sb, and Pb, particularly for Cd (0.98). The high accuracy of Zn and Pb measurements was consistent with the results in Refs. [25,26]. This consistency may be attributed to the pre-treatment of the measured soil through drying, grinding, and sieving, which removed the influence of soil moisture and particle size. In the present study, the Pearson correlation coefficient between pXRF measurement data and ICP measurement data of As was 0.77, while pXRF measurements data and ICP measurement data showed extremely high linear correlation ($R^2=0.999$) in the study of TIAN et al [27], which may be related to the content of As in soil. JIANG [28] noted significant deviations in ICP and pXRF measurements when the element content was low. Furthermore, HU et al [29] suggested that the measurements of As and Ni by pXRF need substantial improvement, with Ni only measurable qualitatively, which could also be influenced by soil types [30].

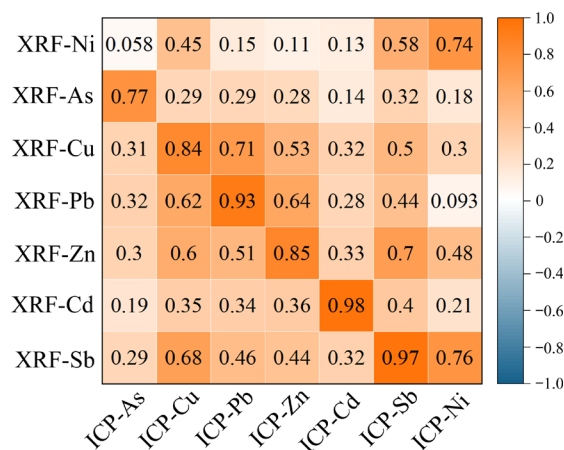


Fig. 2 Pearson correlation coefficient between pXRF and ICP data

The physicochemical properties of all soil samples in the study area were shown in Fig. S1 in Supplementary Materials. The pH values of all soil samples ranged from 8.07 to 8.56 (median: 8.37). The SOM content ranged from 0.72% to 1.11% (median: 0.837%), and the organic matter content

of the soil was relatively low and concentrated. The soil available phosphorus content ranged from 10.12 to 89.82 mg/kg (median: 45.944 mg/kg). The free iron oxides ranged from 8.35 to 14.57 g/kg (median: 11.005 g/kg). The amorphous iron oxides were in the range of 0.45–1.99 g/kg (median: 0.991 g/kg). The crystalline iron oxides were in the range of 7.54–13.72 g/kg (median: 10.016 g/kg).

3.2 Performance of different ML algorithms in predicting soil PTEs

Six different ML algorithms were used to establish the model, and the prediction results were shown in Figs. S2–S7 in Supplementary Materials). The results of the ML algorithm with the best predictive performance for each PTE content were shown in Fig. 3. The dots represented the datasets obtained by pXRF testing against the equivalent expected values calculated by the ML model algorithm. The color change of the scatter indicated the ratio of the actual content of PTEs to the predicted value, which was helpful for visually identifying the overall or partial prediction effect of the model.

The R^2 and RMSE values for each model in predicting seven PTEs were shown in Fig. 4. Both traditional learning and ensemble learning achieved a relatively good level of prediction for Cd and Sb. The stronger predictive performance for Cd and Sb may be due to the excellent agreement between the measurements collected by pXRF and ICP (Fig. 2). Linear regression was used to effectively predict Cd, Sb, and Pb, as they shared a good correlation in the original data. Although the predicted results of Cd were very satisfactory according to R^2 , some prediction results were negative, the linear regression method was not suitable for the direct prediction of Cd concentration. The effect of linear regression on Cu prediction was consistent with the results obtained by XIA et al [23]. Random sample consensus, a linear regression-based approach for handling outliers (see Supplement Materials), demonstrated negligible improvement compared to traditional linear regression (Fig. S2). The decision tree showed good applicability for Ni content prediction, although the results of RMSE were not the smallest (RMSE=19.047), which may be caused by the deviation of the model when dealing with high-value data. The RMSE and R^2 for support vector machine and linear regression had almost the

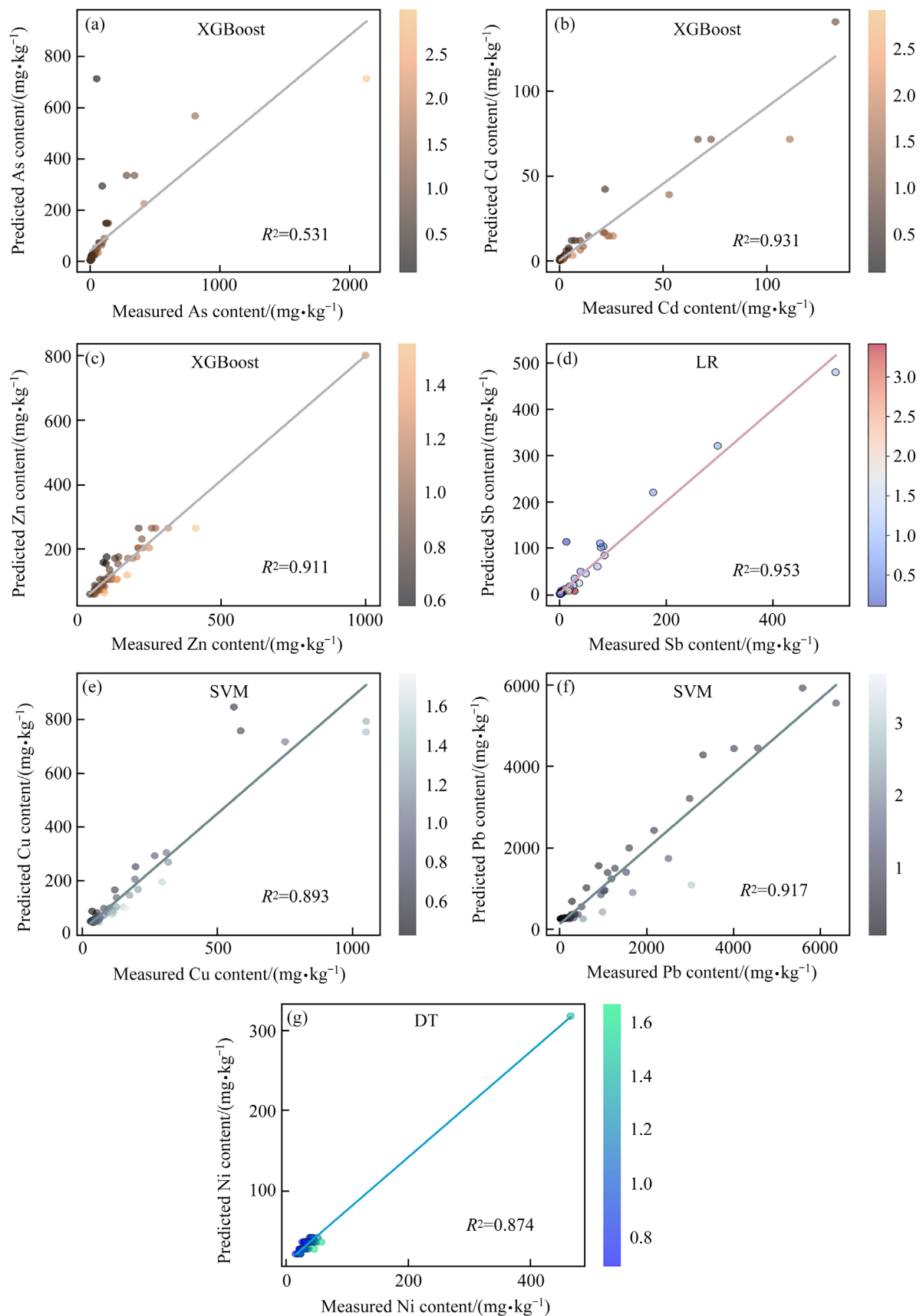


Fig. 3 Contents of PTEs predicted from each pXRF measurement using the best predictive model against measured contents (The line is the fitting line of the scatter and the color of the dot indicates the ratio of the measured value to the predicted value): (a) As; (b) Cd; (c) Zn; (d) Sb; (e) Cu; (f) Pb; (g) Ni

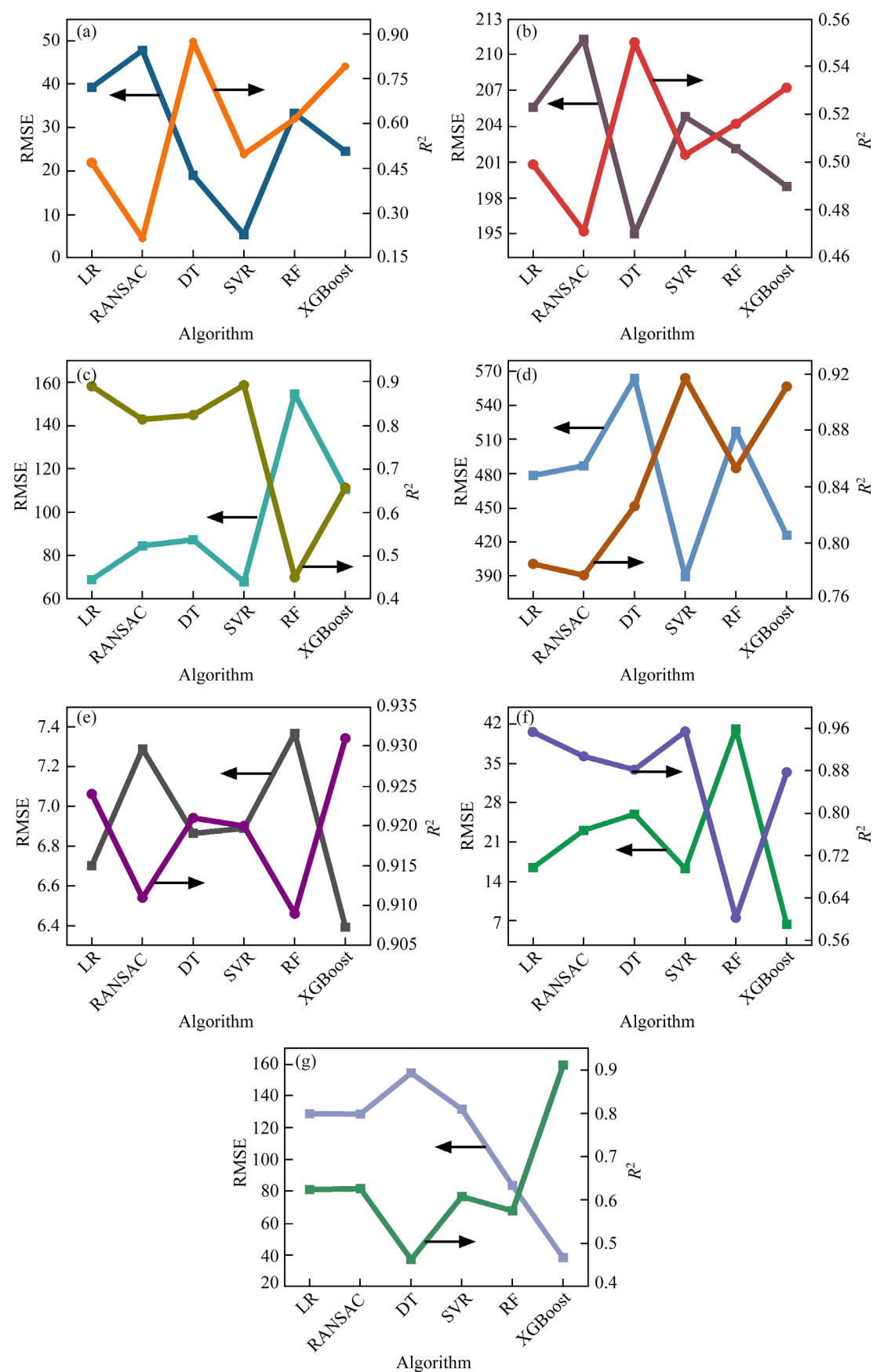


Fig. 4 Comparison of prediction performance of six ML algorithms for Ni (a), As (b), Cu (c), Pb (d), Cd (e), Sb (f) and Zn (g) (Six machine learning algorithms included LR, RANSAC, DT, SVR, RF and XGBoost)

same effect on Cu prediction. However, linear regression yielded a higher deviation than the support vector machine in terms of the maximum ratio between predicted and measured values in the scatter plot (Fig. S3). The support vector machine also performed well in Pb prediction, particularly in the low-value range, and the high RMSE of the prediction results may be due to the high Pb content in the whole site (ICP measured mean value of 697.68 mg/kg). Unfortunately, none of the six algorithms effectively predicted As content. XGBoost outperformed other models in predicting most PTEs contents, with the most significant improvement observed for Zn, reaching a maximum predicted deviation value-to-true value ratio of 1.4. Although Pb prediction was generally accurate across the entire concentration range, there were some instances of significant deviations at lower values.

3.3 Simulation of spatial distribution of PTEs based on optimal ML prediction model

Based on the correction results of different ML algorithms for PTE contents measured by pXRF, the spatial distribution of the predicted values of the best algorithm and measured values for each PTE were shown in Fig. 5. The spatial distribution map of PTEs in the surface soil (0–0.5 m) was mapped using the inverse distance weighting interpolation method in ArcGIS (10.2). Mapping the differences between predicted and measured values was an important step to assess the accuracy of the trained ML models, which helped to evaluate whether the models accurately captured the variation patterns in the target variables across space. Additionally, mapping the predicted contents provided an easy way to visualize and analyze the accuracy of the models. Due to the direct influence of human

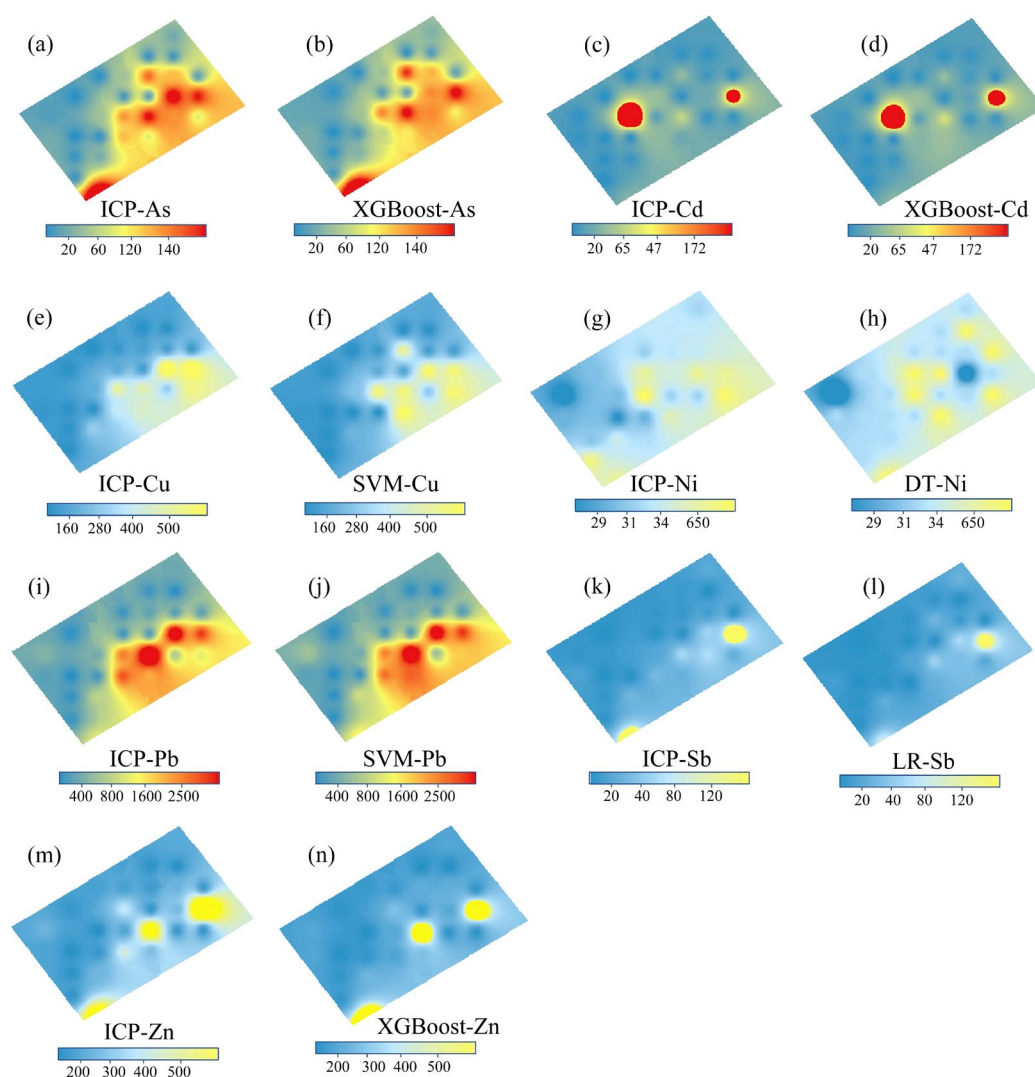


Fig. 5 Comparison of spatial distribution of predicted PTE contents by optimal ML algorithm modeling and measured PTE contents in soil (unit: mg/kg)

activities, the content of PTEs in the surface soil of contaminated sites usually has a high coefficient of variation [31], which was helpful in reflecting whether the models have the capability of capturing both high and low values. The screening value and control value in the reference soil pollution risk control standard for construction land in China (GB 36600—2018) and the US superfund soil screening guidelines were set in sections in the color bar to show the pollution degree of different PTEs in the soil of the plot. Red, yellow, and blue areas represented high, medium, and low contents of PTEs, respectively.

Overall, the predicted PTE contents obtained by the ML models followed the general trends of in the measured values. For As, although the predicted results of XGBoost were unsatisfactory ($R^2=0.531$), the distribution trend was similar. The main deviation observed in XGBoost's prediction was the underestimation of As content at individual high and median points, which likely contributed to the significant deviation in the prediction results (Figs. 5(a, b)). The prediction results for Cd and Sb showed good predictability in terms of the regression coefficients ($R^2=0.931$ and 0.953) and RMSE (6.393 and 16.449). The spatial characteristics of the measured and predicted values of Cd and Sb were not significantly different (Fig. 5). The deviations in Pb content prediction made by SVR were relatively small at various points on the surface soil, and the deviation of some low values may appear in the prediction of deep soil (Fig. S3(e)). In terms of Cu content prediction, SVM yielded slightly lower predicted values compared to the actual values, and there was a significant deviation observed only at one point in the topsoil.

3.4 Influence of soil environmental factors on model prediction accuracy

The topsoil was seriously polluted by Pb, Cd, and As (Fig. 5). Additionally, none of the six ML algorithms achieved satisfactory results in the As prediction when only using pXRF data for prediction. Therefore, the physicochemical properties including organic matter content, pH, available phosphorus content, free iron oxide, amorphous iron oxide, crystalline iron oxide content, and pXRF measurements were further used as model correction factors. XGBoost showed

excellent prediction performance for most PTEs. Compared to other ML algorithms, the XGB algorithm also had the advantage of interpreting the complex soil data, so it was chosen as the algorithm for the multi-input model. SHAP was used to rank and show the importance of various environmental factors in the prediction process, so as to explore the impact of site environmental factors on the prediction process (Fig. 6).

Compared with the single input model ($R^2=0.531$) (Fig. 3(a)), the XGBoost multi-input model significantly improved the prediction efficiency of As ($R^2=0.835$) (Fig. 6(b)). In the prediction process of As, only pH and amorphous iron oxide content played an important role. Regarding the process of Cd prediction, pH, organic matter content, free iron oxide, and available phosphorus content had a certain role. In the process of Pb prediction, six soil physicochemical properties were involved, with organic matter playing the most crucial role. The overall color distribution of scattered points indicated higher prediction accuracy of the model in the low value of Pb. Whether for Pb, Cd or As, pH and organic matter played an important role in predicting. Numerous studies [31–33] showed that pH and soil organic matter had a dominant influence on the bioavailability and mobility of contaminants. Organic matter in soil easily adsorbs soil PTEs. Previous research [32] showed that lighter carbon and hydrogen elements in organic matter may dilute Pb content in the soil during pXRF measurement. However and LEMKE the dilution effect of organic matter on PTEs also exists in elemental dependence. For example, RAVANSARI [33] showed that the pXRF content of Pb was in good agreement with the theoretical dilution line of organic matter; on the contrary, the pXRF content of As was always lower than the theoretical dilution line of organic matter. This also indicated that the interference of organic matter in the pXRF measurement process cannot be adequately described and corrected by simple linear equations.

As had a high affinity for metal oxides in soil, especially iron oxides [34]. Amorphous iron oxides can adsorb As more strongly than free iron oxides, such as goethite and magnetite, as well as crystalline iron oxides [35]. However, this adsorption largely affects the morphology of As present in the soil,

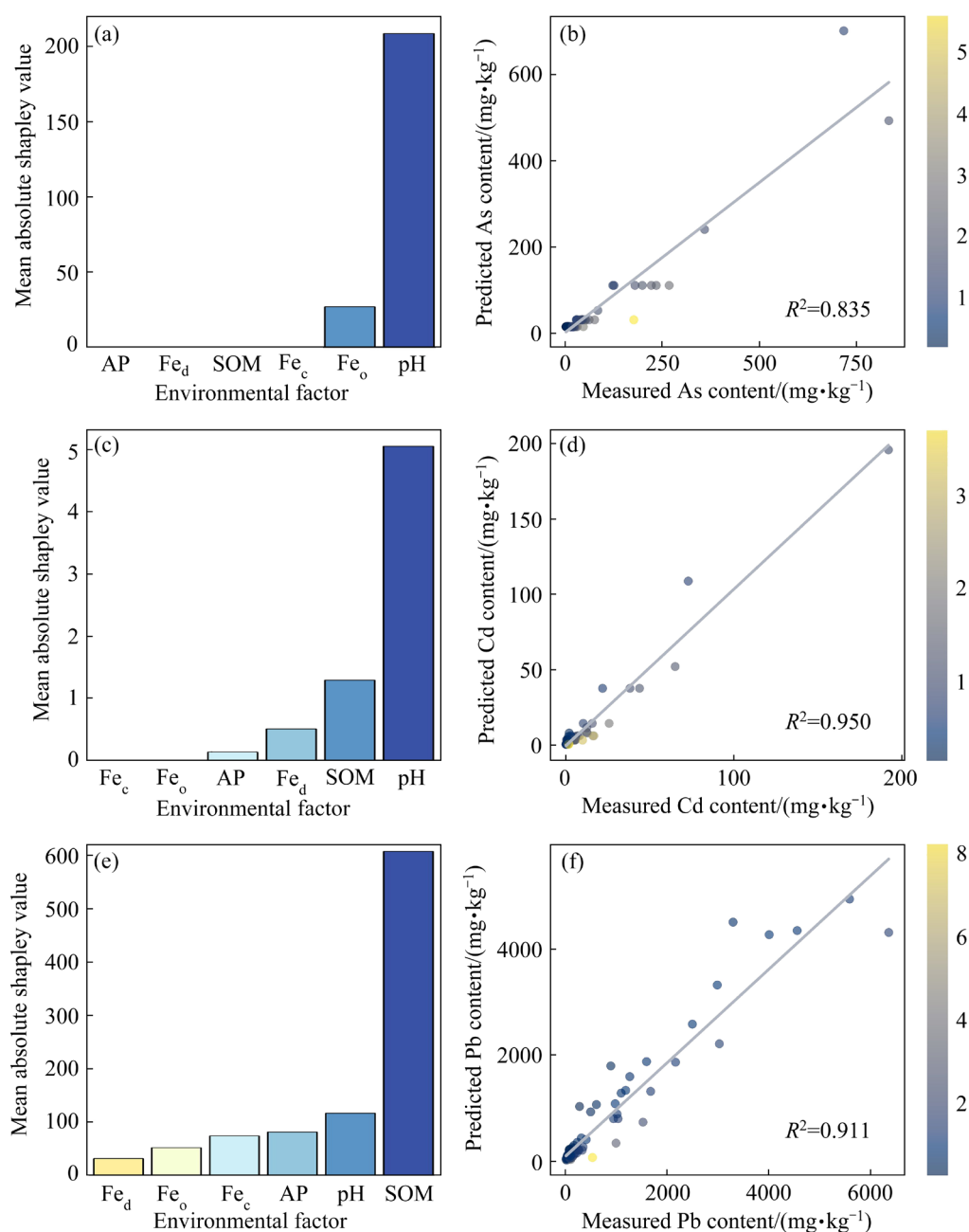


Fig. 6 Mean absolute Shapley values showing average impact on model output magnitude for all descriptors in XGBoost model (a, c, e), and predicted PTE contents using XGBoost model based on pXRF measurement and environmental characteristic values versus measured values (b, d, f): (a, b) As; (c, d) Cd; (e, f) Pb

which can interfere with pXRF measurements. Organic matter played a crucial role in controlling the transformation of PTE fugitive forms in soils, and high organic matter content tends to limit the migration flow of PTEs [36]. In the soil, the active materials include iron minerals and organic matter, which usually exist as iron-organic associations. The adsorption of Cd can be strongly influenced by iron-organic associations [37]. Organic matter, clay minerals, and iron oxide were the three adsorbents with strong sorption of Pb in the soil [38].

Pb in the soil can form ternary complexes (Fe-OC-Pb, Fe-Pb-OC) with organic matter and iron oxides [39], resulting in the presence of Pb in the soil in the organic-bound and Fe-Mn oxide-bound states. Phosphates in soils were able to transform with PTEs ions from a readily transportable water-soluble state to a stable precipitate, e.g., Pb and phosphate form the most stable environmental soil lead, lead phosphate (pyromorphite mineral family ($Pb_5(PO_4)_3X$; X= halide or hydroxide)) [40]. Similarly, Cd forms

cadmium phosphate precipitates with phosphate.

The occurrence forms of PTEs in soil were highly susceptible to the influence of soil pH, which was widely considered to be one of the most important geochemical parameters affecting the present form of PTEs in soil [11,41]. For example, the mobility of Cd and Pb increases with the increase of pH, and their mobility is much higher in acidic soils than in alkaline and neutral soils [42]. WIECZOREK et al [43] found that the solubility of Cd and Pb was reduced in soils with low soil pH but high organic matter content. Conversely, the mobility of As might increase in alkaline environments due to competition for anion sites [44].

3.5 Adapted ecological risk assessment of PTEs in profile soils

The potential ecological risk of PTEs was evaluated using an adapted method that considered six environmental characteristics influencing the toxicity and mobility of these metals. Based on these characteristics, the ecological risk of the soil at the site was calculated. The adapted potential ecological risk index (ARI) was calculated by Formulas (1)–(4) based on the PTE contents of the control sites around the contaminated site and the background values of PTEs in Chinese soils (Table 2), and compared with the grading criteria of Hakanson ecological risk index (Tables S1 and S2 in Supplementary Materials). Pb, Cd, and As have different degrees of potential ecological risk at the site. Except for Cd, the mean values of the potential ecological risk indices for other six PTEs were below the moderate risk threshold ($E=40$). Among

all soil samples, Cd was identified as the main contributor to the high risk, as 25% soil samples had a high potential ecological risk for Cd, and 7% soil samples had a very high potential ecological risk for Cd. In contrast, only a low percentage of soil samples had a high ecological risk for As (1% soil samples), while 8% soil samples had a potential ecological risk for Pb. Other PTEs, including Sb, Ni, Cu, and Zn, showed low ecological risk.

Table 2 Descriptive statistics for adapted individual potential risk value (AE)

Metal	Max	Min	Mean
Pb	51.5	0.16	2.7
Cd	106	0.2	4.67
As	328	1.22	39.3
Ni	1.8	0.03	0.14
Cu	0.28	0.01	0.04
Zn	19.1	0.24	1.54
Sb	17.8	0.04	0.52

Referring to the grading criteria of Hakanson potential ecological risk index in Table S1, the comprehensive potential ecological risk was divided into five levels, and the spatial distribution of ecological risk at different depths was mapped by ArcGIS, with different levels of red and yellow representing the existence of different levels of comprehensive potential ecological risk (Fig. 7). The ecological risks caused by PTEs at the site exist in the entire depth of investigation, and the potential ecological risks at depths below 2.5 m

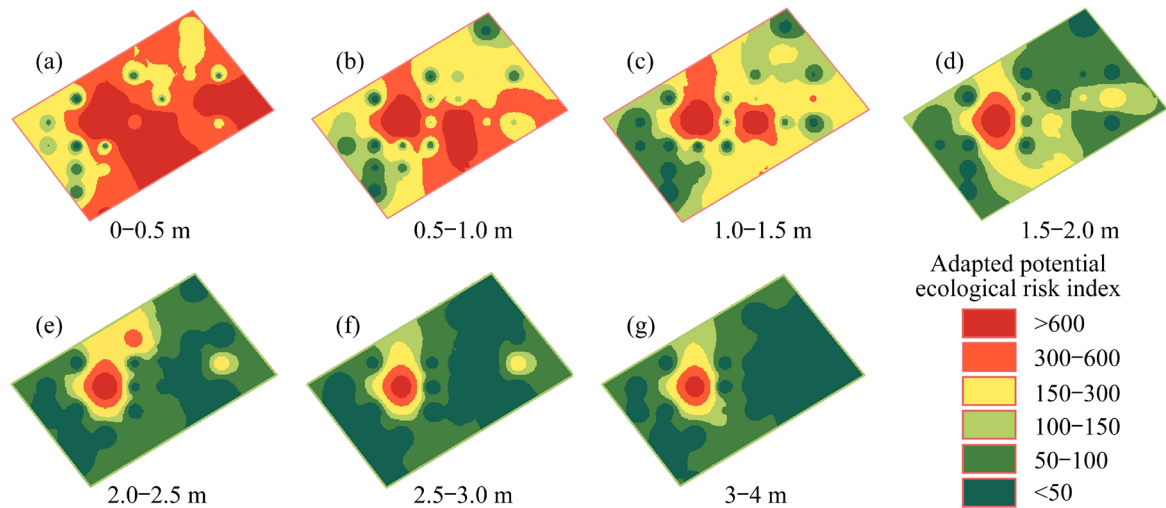


Fig. 7 Spatial distribution of adapted potential ecological risk index at different depths

mainly concentrate at a particular sampling site, which may result from the presence of high content of Cd, the high mobility of Cd causing the entire depth range to be affected [43], its high toxicity (toxicity efficient $T_{cd}=30$) resulting in a high ecological risk, and the potential ecological risks presented in the deep soil may further affect the soil ground-water [45,46]. The spatial distribution of ecological risks corresponds closely to the presence of elevated PTEs concentrations (Pb, As, and Cd).

The distribution of ecological risks present in the surface soil was consistent with the distribution of Pb, As, and Cd in the surface soil. Localized areas surrounding the study site exhibited pronounced influences attributable to anthropogenic processes, with major contributions stemming from industrial activities. The high potential ecological risks in the surface soil were mainly concentrated near the production area, and linked to leakage of pollutants in the plant during the industrial production process. The ecological risks of Pb, Cd, and As in the study site require urgent attention in subsequent site remediation and control.

The adapted risk assessment considered a wider range of factors, effectively uncovering the intricate relationship between pollutants and environmental features. This holistic approach offered significant advantages over traditional ecological risk assessment that focused solely on the content of PTEs in the environment. Traditional ecological risk assessment focused on the content of PTEs in the environment, which may result in an overestimation of potential risks for a given site. By considering various environmental factors, adapted potential ecological risk assessment provides a more nuanced perspective on ecological risk. Therefore, potential ecological risk assessment based on environmental factors offers a more comprehensive and accurate approach to assessing ecological risk.

4 Conclusions

(1) DT had the best performance in predicting Ni. SVM had good predictions for Cu and Pb. XGBoost performed well for most PTEs, except As. However, by incorporating environmental factors as model correction factors, the prediction of As was significantly improved.

(2) The determination process of Pb, Cd, and

As was affected by pH and soil organic matter content. Additionally, pH and soil organic matter content played key roles in developing the generalized prediction model.

(3) The spatial distribution of ecological risk was consistent with that of PTEs (Pb, Cd, and As). Although the adapted potential ecological risk index reduced the ecological risk to some extent compared with the Hakanson ecological risk index, the single factor ecological risk index still indicated that there were different degrees of ecological risk for Pb, Cd, and As at the site, with Cd being the primary contributor. Therefore, it is necessary to focus on the remediation of Pb, Cd, and As in the subsequent site remediation and pollution control.

CRedit authorship contribution statement

Sheng-guo XUE: Conceptualization, Methodology, Investigation, Validation, Data curation, Formal analysis, Writing – Original draft, Writing – Review & editing; **Jing-pei FENG:** Investigation; **Wen-shun KE:** Conceptualization, Methodology, Validation, Investigation, Supervision, Writing – Review & editing; **Mu LI:** Writing – Review & editing; **Kun-yan QIU** and **Chu-xuan LI:** Investigation; **Chuan WU:** Resources; **Lin GUO:** Conceptualization, Methodology, Supervision, Resources, Writing – Review & editing, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was financially supported from the National Key Research and Development Program of China (No. 2019YFC1803601), the Fundamental Research Funds for the Central Universities of Central South University, China (No. 2023ZZTS0801), the Postgraduate Innovative Project of Central South University, China (No. 2023XQLH068), and the Postgraduate Scientific Research Innovation Project of Hunan Province, China (No. QL20230054).

Supplementary Materials

Supplementary Materials in this paper can be found at: http://tnmsc.csu.edu.cn/download/24-p3054-2023-0274-Supplementary_Materials.pdf.

References

- [1] HU Ying-yan, LI Min, WU Wei-guo, KE yong, LIU Lu-jing, WANG Xue-liang. Life cycle assessment for waste acid treatment in zinc smelting [J]. Transactions of Nonferrous Metals Society of China, 2022, 32(11): 3822–3834.
- [2] ZENG Jia-qing, LI Chu-xuan, WANG Jin-ting, TANG Lu, WU Chuan, XUE Sheng-guo. Pollution simulation and remediation strategy of a zinc smelting site based on multi-source information [J]. Journal of Hazardous Materials, 2022, 433: 128774.
- [3] ETTLER V. Soil contamination near non-ferrous metal smelters: A review [J]. Applied Geochemistry, 2016, 64: 56–74.
- [4] ZENG Jia-qing, TABELIN C B, GAO Wen-yan, TANG Lu, LUO Xing-hua, KE Wen-shun, JIANG Jun, XUE Sheng-guo. Heterogeneous distributions of heavy metals in the soil-groundwater system empowers the knowledge of the pollution migration at a smelting site [J]. Chemical Engineering Journal, 2023, 454: 140307.
- [5] JIANG Zhi-chao, GUO Zhao-hui, PENG Chi, LIU Xu, ZHOU Zi-ruo, XIAO Xi-yuan. Heavy metals in soils around non-ferrous smelteries in China: Status, health risks and control measures [J]. Environmental Pollution, 2021, 282: 117038.
- [6] WEINDORF D C, BAKR N, ZHU Y D. Advances in portable X-ray fluorescence (PXRF) for environmental, pedological, and agronomic applications [J]. Advances in Agronomy, 2014, 128: 1–45.
- [7] DURANCE P, JOWITT S M, BUSH K. An assessment of portable X-ray fluorescence spectroscopy in mineral exploration, Kurnalpi terrane, Eastern goldfields Superterrane, Western Australia [J]. Applied Earth Science, 2014, 123(3): 150–163.
- [8] DASGUPTA S, CHAKRABORTY S, WEINDORF D C, LI B, SILVA S H G, BHATTACHARYYA K. Influence of auxiliary soil variables to improve PXRF-based soil fertility evaluation in India [J]. Geoderma Regional, 2022, 30: e00557.
- [9] KE Wen-shun, LI Chu-xuan, ZHU Feng, LUO Xing-hua, FENG Jing-pei, LI Xue, JIANG Yi-fan, WU Chuan, HARTLEY W, XUE Sheng-guo. Effect of potentially toxic elements on soil multifunctionality at a lead smelting site [J]. Journal of Hazardous Materials, 2023, 454: 131525.
- [10] YAO Li-wei, MIN Xiao-bo, XU Hui, KE Yong, WANG Yun-yan, LIN Zhang, LIANG Yan-jie, LIU De-gang, XU Qiu-jing, HE Yu-yang. Physicochemical and environmental properties of arsenic sulfide sludge from copper and lead-zinc smelter [J]. Transactions of Nonferrous Metals Society of China, 2020, 30(7): 1943–1955.
- [11] ZHONG Xi, CHEN Zi-wu, LI Ya-ying, DING Keng-bo, LIU Wen-shen, LIU Ye, YUAN Yong-qiang, ZHANG Miao-yue, BAKER A J M, YANG Wen-jun, FEI Ying-heng, WANG Yu-jie, CHAO Yuan-qing, QIU Rong-liang. Factors influencing heavy metal availability and risk assessment of soils at typical metal mines in Eastern China [J]. Journal of Hazardous Materials, 2020, 400: 123289.
- [12] KE Wen-shun, LI Chu-xuan, ZHU Feng, LUO Xing-hua, LI Xue, WU Chuan, HARTLEY W, XUE Sheng-guo. The assembly process and co-occurrence patterns of soil microbial communities at a lead smelting site [J]. Science of the Total Environment, 2023, 894: 164932.
- [13] ZENG Jia-qing, LUO Xing-hua, CHENG Yi-zhi, KE Wen-shun, HARTLEY W, LI Chu-xuan, JIANG Jun, ZHU Feng, XUE Sheng-guo. Spatial distribution of toxic metal(loid)s at an abandoned zinc smelting site, Southern China [J]. Journal of Hazardous Materials, 2022, 425: 127970.
- [14] CAPORALE A G, ADAMO P, CAPOZZI F, LANGELLA G, TERRIBILE F, VINGIANI S. Monitoring metal pollution in soils using portable-XRF and conventional laboratory-based techniques: Evaluation of the performance and limitations according to metal properties and sources [J]. Science of the Total Environment, 2018, 643: 516–526.
- [15] WANG Zhen-xing, YU Yun-jun, YE Tian-tian, FEI Jiang-chi, SONG Xin-yu, PENG Jian-wei, ZHOU Yao-yu, WU Hong-hua. Distribution characteristics and environmental risk assessment following metal(loid)s pollution incidents at Southwest China mining site [J]. Transactions of Nonferrous Metals Society of China, 2022, 32(12): 4062–4075.
- [16] GASPARATOS D. Sequestration of heavy metals from soil with Fe–Mn concretions and nodules [J]. Environmental Chemistry Letters, 2013, 11(1): 1–9.
- [17] ZHANG Jian, JIANG Yin-kun, DING Cheng-yu, WANG Sheng-sen, ZHAO Chen-hao, YIN Wei-qin, WANG Bing, YANG Rui-dong, WANG Xiao-zhi. Remediation of lead and cadmium co-contaminated mining soil by phosphate-functionalized biochar: Performance, mechanism, and microbial response [J]. Chemosphere, 2023, 334: 138938.
- [18] XUE Sheng-guo, KORNA R, FAN Jia-rong, KE Wen-shun, LOU Wei, WANG Jin-ting, ZHU Feng. Spatial distribution, environmental risks, and sources of potentially toxic elements in soils from a typical abandoned antimony smelting site [J]. Journal of Environmental Sciences, 2023, 127: 780–790.
- [19] YANG Jie-jie, GUO Zi-wen, JIANG Lu-hua, SARKODIE E K, LI Ke-wei, SHI Jia-xin, DENG Yan, ZHANG Zi-cheng, LIU Hong-wei, LIANG Yi-li, YIN Hua-qun, LIU Xue-duan. Cadmium, lead and arsenic contamination in an abandoned nonferrous metal smelting site in Southern China: Chemical speciation and mobility [J]. Ecotoxicology and Environmental Safety, 2022, 239: 113617.
- [20] ELITH J, LEATHWICK J R, HASTIE T. A working guide to boosted regression trees [J]. Journal of Animal Ecology, 2008, 77(4): 802–813.
- [21] ANDRADE R, FARIA W M, SILVA S H G, CHAKRABORTY S, WEINDORF D C, MESQUITA L F, GUILHERME L R G, CURI N. Prediction of soil fertility via portable X-ray fluorescence (pXRF) spectrometry and soil texture in the Brazilian Coastal Plains [J]. Geoderma, 2020, 357: 113960.
- [22] YANG Hong-rui, HUANG Kuan, ZHANG Kai, WENG Qin, ZHANG Hui-chun, WANG Fei-er. Predicting heavy metal

- adsorption on soil with machine learning and mapping global distribution of soil adsorption capacities [J]. *Environmental Science & Technology*, 2021, 55(20): 14316–14328.
- [23] XIA Fei-yang, FAN Ting-ting, CHEN Yun, DING Da, WEI Jing, JIANG Deng-deng, DENG Shao-po. Prediction of heavy metal concentrations in contaminated sites from portable X-ray fluorescence spectrometer data using machine learning [J]. *Processes*, 2022, 10(3): 536.
- [24] QU Ming-kai, GUANG Xu, LIU Hong-bo, ZHAO Yong-cun, HUANG Biao. Additional sampling using in-situ portable X-ray fluorescence (PXRF) for rapid and high-precision investigation of soil heavy metals at a regional scale [J]. *Environmental Pollution*, 2022, 292: 118324.
- [25] MCCOMB J Q, ROGERS C, HAN F X, TCHOUNWOU P B. Rapid screening of heavy metals and trace elements in environmental samples using portable X-ray fluorescence spectrometer: A comparative study [J]. *Water, Air, & Soil Pollution*, 2014, 225(12): 2169.
- [26] BERNICK M B, KALNICKY D J, PRINCE G, SINGHVI R. Results of field-portable X-ray fluorescence analysis of metal contaminants in soil and sediment [J]. *Journal of Hazardous Materials*, 1995, 43: 101–110.
- [27] TIAN Kang, HUANG Biao, XING Zhe, HU Wen-you. In situ investigation of heavy metals at trace concentrations in greenhouse soils via portable X-ray fluorescence spectroscopy [J]. *Environmental Science and Pollution Research*, 2018, 25(11): 11011–11022.
- [28] JIANG Min. Application of portable X-ray fluorescence (pXRF) for heavy metal analysis of soils in crop fields near abandoned mine sites [J]. *Environmental Geochemistry and Health*, 2010, 32(3): 207–216.
- [29] HU Bi-feng, CHEN Song-chao, HU Jie, XIA Fang, XU Jun-feng, LI Yan, SHI Zhou. Application of portable XRF and VNIR sensors for rapid assessment of soil heavy metal pollution [J]. *PLoS One*, 2017, 12(2): e0172438.
- [30] FRAHM E, DOONAN R C P. The technological versus methodological revolution of portable XRF in archaeology [J]. *Journal of Archaeological Science*, 2013, 40(2): 1425–1434.
- [31] KE Wen-shun, ZENG Jia-qing, ZHU Feng, LUO Xing-hua, FENG Jing-pei, HE Jin, XUE Sheng-guo. Geochemical partitioning and spatial distribution of heavy metals in soils contaminated by lead smelting [J]. *Environmental Pollution*, 2022, 307: 119486.
- [32] SHUTTLEWORTH E L, EVANS M G, HUTCHINSON S M, ROTHWELL J J. Assessment of lead contamination in peatlands using field portable XRF [J]. *Water, Air, & Soil Pollution*, 2014, 225(2): 1844.
- [33] RAVANSARI R, LEMKE L D. Portable X-ray fluorescence trace metal measurement in organic rich soils: pXRF response as a function of organic matter fraction [J]. *Geoderma*, 2018, 319: 175–184.
- [34] XU Xiao-wei, CHEN Chuan, WANG Peng, KRETZ-SCHMAR R, ZHAO Fang-jie. Control of arsenic mobilization in paddy soils by manganese and iron oxides [J]. *Environmental Pollution*, 2017, 231: 37–47.
- [35] DIXIT S, HERING J G. Comparison of arsenic(V) and arsenic(III) sorption onto iron oxide minerals: Implication for arsenic mobility [J]. *Environmental Science & Technology*, 2003, 37(18): 4182–4189.
- [36] WAN Dan, ZHANG Ni-chen, CHEN Wen-li, CAI Peng, ZHENG Li-rong, HUANG Qiao-yun. Organic matter facilitates the binding of Pb to iron oxides in a subtropical contaminated soil [J]. *Environmental Science and Pollution Research*, 2018, 25(32): 32130–32139.
- [37] MAO Yi-jie, FAN Wei-guo, YAN Ya-xin, XIANG Wu, HU Sheng-hong, YAN Sen. Cd adsorption by iron-organic associations: Implications for Cd mobility and fate in natural and contaminated environments [J]. *Bulletin of Environmental Contamination and Toxicology*, 2021, 106(1): 109–114.
- [38] CAO X D, MA L Q. Effects of compost and phosphate on plant arsenic accumulation from soils near pressure-treated wood [J]. *Environmental Pollution*, 2004, 132(3): 435–442.
- [39] XIONG Juan, KOOPAL L K, WENG Li-ping, WANG Ming-xia, TAN Wen-feng. Effect of soil fulvic and humic acid on binding of Pb to goethite–water interface: Linear additivity and volume fractions of HS in the Stern layer [J]. *Journal of Colloid and Interface Science*, 2015, 457: 121–130.
- [40] RUBY M V, DAVIS A, NICHOLSON A. In situ formation of lead phosphates in soils as a method to immobilize lead [J]. *Environmental Science & Technology*, 1994, 28(4): 646–654.
- [41] MSAKY J J, CALVET R. Adsorption behavior of copper and zinc in soils: Influence of pH on adsorption characteristics [J]. *Soil Science*, 1990, 150(2): 513–522.
- [42] DONG De-ming, ZHAO Xing-min, HUA Xiu-yi, LIU Jin-fu, GAO Ming. Investigation of the potential mobility of Pb, Cd and Cr(VI) from moderately contaminated farmland soil to groundwater in Northeast, China [J]. *Journal of Hazardous Materials*, 2009, 162(2/3): 1261–1268.
- [43] WIECZOREK J, BARAN A, URBANSKI K, MAZUREK R, KLIMOWICZ-PAWLAS A. Assessment of the pollution and ecological risk of lead and cadmium in soils [J]. *Environmental Geochemistry and Health*, 2018, 40(6): 2325–2342.
- [44] MASSCHELEYN P H, DELAUNE R D, PATRICK W H Jr. Effect of redox potential and pH on arsenic speciation and solubility in a contaminated soil [J]. *Environmental Science & Technology*, 1991, 25: 1414–1419.
- [45] TANG Lu, LIU Jie, ZENG Jia-qing, LUO Xing-hua, KE Wen-shun, LI Chu-xuan, GAO Wen-yan, JIANG Jun, XUE Sheng-guo. Anthropogenic processes drive heterogeneous distributions of toxic elements in shallow groundwater around a smelting site [J]. *Journal of Hazardous Materials*, 2023, 453: 131377.
- [46] XUE Sheng-guo, KE Wen-shun, ZENG Jia-qing, TABELIN C B, XIE Yi, TANG Lu, XIANG Chao, JIANG Jun. Pollution prediction for heavy metals in soil-groundwater systems at smelting sites [J]. *Chemical Engineering Journal*, 2023, 473: 145499.

基于机器学习的铅冶炼场地污染快速识别与风险评估

薛生国¹, 冯静培¹, 可文舜¹, 李 幕¹, 邱坤艳², 李楚璇¹, 吴 川¹, 郭 林³

1. 中南大学 冶金与环境学院, 长沙 4100083;
2. 河南省土壤重金属污染监测与修复重点实验室, 济源 454650;
3. 河南省地质研究院, 郑州 450000

摘 要: 以便携式 X 射线荧光光谱仪(pXRF)测量的 207 个土壤样品的重金属含量和 6 种环境因素作为模型修正系数, 建立了 7 种重金属的通用预测模型。为了评估冶炼场地存在的潜在生态风险, 使用 XGBoost 算法拟合重金属含量和环境特征之间的关系, 建立了基于环境因素的潜在生态风险指数。结果表明, 通用预测模型对铅(拟合系数 $R^2=0.911$)、镉($R^2=0.950$)和砷($R^2=0.835$)均具有极高的预测精度; 表层土壤重金属的潜在生态风险较高, 部分点位因 Cd 的高迁移性在不同深度均有较高的潜在生态风险; 机器学习显著提高了 pXRF 重金属测量结果的准确性, 识别了影响测量过程的关键环境因素。基于改进的潜在生态风险评价, 该铅冶炼场地铅、镉和砷的生态风险较高, 应重点考虑对其进行修复。

关键词: 冶炼场地; 潜在有害元素; X 射线荧光光谱; 潜在生态风险; 机器学习

(Edited by Wei-ping CHEN)