ACS Partner Journal

ENVIRONMENT & HEALTH

Open Access

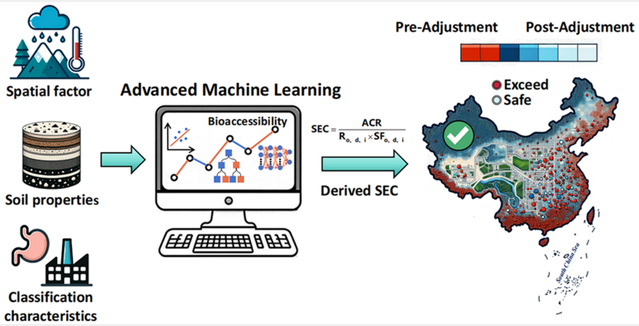
Predicting the Bioaccessibility of Soil Cd, Pb, and As with Advanced Machine Learning for Continental-Scale Soil Environmental Criteria Determination in China

Kunting Xie, Jiajun Ou, Minghao He, Weijie Peng, and Yong Yuan*

Cite This: *Environ. Health* 2024, 2, 631−641

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Investigating the bioaccessibility of harmful inorganic elements in soil is crucial for understanding their behavior in the environment and accurately assessing the environmental risks associated with soil. Traditional batch experimental methods and linear models, however, are time-consuming and often fall short in precisely quantifying bioaccessibility. In this study, using 937 data points gathered from 56 journal articles, we developed machine learning models for three harmful inorganic elements, namely, Cd, Pb, and As. After thorough analysis, the model optimized through a boosting ensemble strategy demonstrated the best performance, with an average R^2 of 0.95 and an RMSE of 0.25. We further employed SHAP values in conjunction with quantitative analysis to identify the key features that influence bioaccessibility. By utilizing the developed integrated models, we carried out predictions for 3002 data points across China, clarifying the bioaccessibility of cadmium (Cd), lead (Pb), and arsenic (As) in the soils of various sites and constructed a comprehensive spatial distribution map of China using the inverse distance weighting (IDW) interpolation method. Based on these findings, we further derived the soil environmental standards for metallurgical sites in China. Our observations from the collected data indicate a reduction in the number of sites exceeding the standard levels for Cd, Pb, and As in mining/smelting sites from 5, 58, and 14 to 1, 24, and 7, respectively. This research offers a precise and scientific approach for cross-regional risk assessment at the continental scale and lays a solid foundation for soil environmental management.

KEYWORDS: *machine learning, soil environmental criteria, potentially harmful elements, bioaccessibility, sites*

1. INTRODUCTION

Soil pollution, primarily caused by mining activities and uncontrolled industrial emissions, has become a significant global challenge.[1−3] China's first pollution census indicated that 16% of soil samples and approximately one-third of metallurgical and industrial sites exceeded pollution standards, with Cd, Pb, and As being the primary inorganic contaminants.[4] In response, China introduced the "Soil Pollution Prevention and Control Action Plan," aiming for 95% of contaminated land to be safely utilized by 2030.[5] Soil environmental standards play an indispensable role in the assessment of potential exposure risks and in the determination of the need for more in-depth site investigation.[6] Although the Chinese government has issued a unified national standard for site soils based on health risks for soil pollution screening, given China's vast land area and diverse soil types, a uniform standard may not be adequate to scientifically and comprehensively assess exposure risks across different regions.[7]

Methodologies for human health risk assessment must be developed to establish site exposure criteria (SECs) based on health risks.[8,9] Notably, incidental ingestion is an important route for the intake of metals from soil.[10,11] It is imperative to ascertain the bioavailability of potentially harmful elements (PHE) in contaminated soil matrices to accurately determine human exposure. Currently, oral bioavailability is gauged by comparing metal accumulations in animal tissues or urine, with these animals being exposed to soluble reference compounds such as sodium arsenate (NaH_2AsO_4), lead acetate ($Pb(AC)_2$), or cadmium chloride ($CdCl_2$).[12−14] Owing to the substantial costs and ethical dilemmas linked with animal testing, the assessment of bioaccessibility—the proportion of heavy metals extracted in vitro from gastric simulations relative to their total content—has been widely explored as an alternative for appraising PHE bioavailability. This approach not only

Received: February 15, 2024
Revised: May 22, 2024
Accepted: May 22, 2024
Published: June 6, 2024

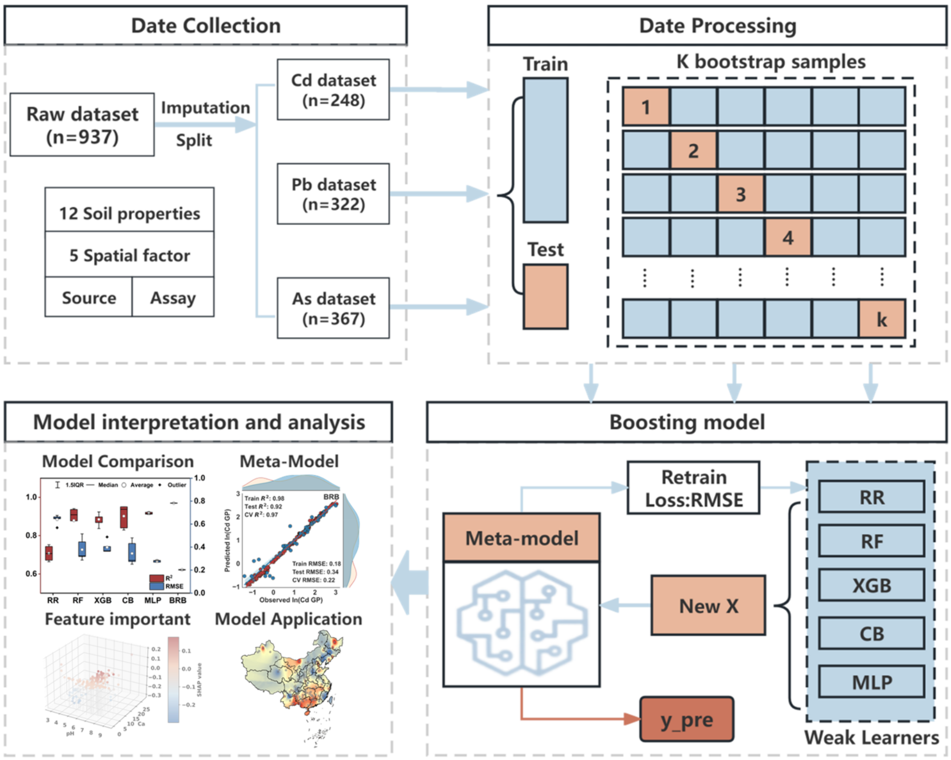


Figure 1. Schematic diagram of the model building process.

circumvents the limitations of animal-based tests but has also been corroborated by them and extensively applied in research.[15−18] While several research groups have evaluated the health risks of site pollutants based on oral bioavailability/bioaccessibility, they predominantly relied on limited, site-specific pollutant data.[19,20] To date, the absence of nationwide investigations means that regional variations in PHE bioaccessibility on a continental level and their subsequent implications for health risk evaluations are yet to be fully understood.

The distinct properties of soil, such as pH level, cation exchange capacity, particle size distribution, and nutrient content, lead to significant variations in the bioavailability of PHE in different soil environments.[21,22] Traditionally employed methods for determining PHEs' bioaccessibility in site soil are noted for their inefficiency and extensive time requirements. An alternative approach is to utilize traditional multivariate linear regression models to estimate the bioaccessibility of PHE in site soils.[23,24] However, these traditional models have a narrow application range, as batch experiments are a prerequisite for obtaining the bioaccessibility of soil PHE before modeling.[25] Furthermore, due to the heterogeneity of soils, it is challenging to use these traditional large-scale models to quantify the bioaccessibility of soil PHE, as the experimental parameters need to be predetermined before developing models for specific types of site soils. Given these shortcomings, machine learning models have been widely applied in the field of environmental science as powerful tools for discovering complex relationships, owing to their low cost, high accuracy, and robustness.[26−28] Although the bioaccessibility of PHE in site soils is a key research area, related studies are still relatively limited. Several researchers have already introduced machine learning methods to predict the bioaccessibility of PHE. For instance, Xie et al. utilized a random forest (RF) model to predict the bioavailability of heavy metals in soil using samples from 12 metallurgical sites.[29] In contrast, Zhang et al. employed conditional inference trees (CIT) and RF models to explore the primary factors influencing heavy metal bioavailability in paddy soils in the karst regions of the northern part of Guangxi, China.[30] However, due to limitations in the data, a lack of diversity in soil types, and instability in model performance, the applicability of the models developed in these studies is limited. To overcome these barriers, there is a need for more holistic and representative data aggregation, coupled with further refinement and authentication of models, catering to the need for predictive models that can be applied across diverse environments and scenarios.

Therefore, in this study, data-driven methods are introduced to explore new approaches for improving the soil environmental standards of regional sites. To assess the applicability of machine learning in the study of site soil environmental standards, we framed a comprehensive data set for the bioaccessibility of three different PHE in various soils, utilizing 937 data points. This data set was used to develop machine learning (ML) models based on five different ML algorithms, and the best-performing parameters were selectively integrated to construct more stable predictive models for soil PHE bioaccessibility. The models that performed best were used to identify the key factors affecting soil PHE. To evaluate the applicability of the ML models in determining SEC, we applied the integrated model to predict the bioaccessibility of soil PHE in 3002 samples from 31 provinces in China. Based on the model prediction results, we derived typical SEC, offering a comprehensive understanding of soil PHE bioaccessibility across different regions in China.

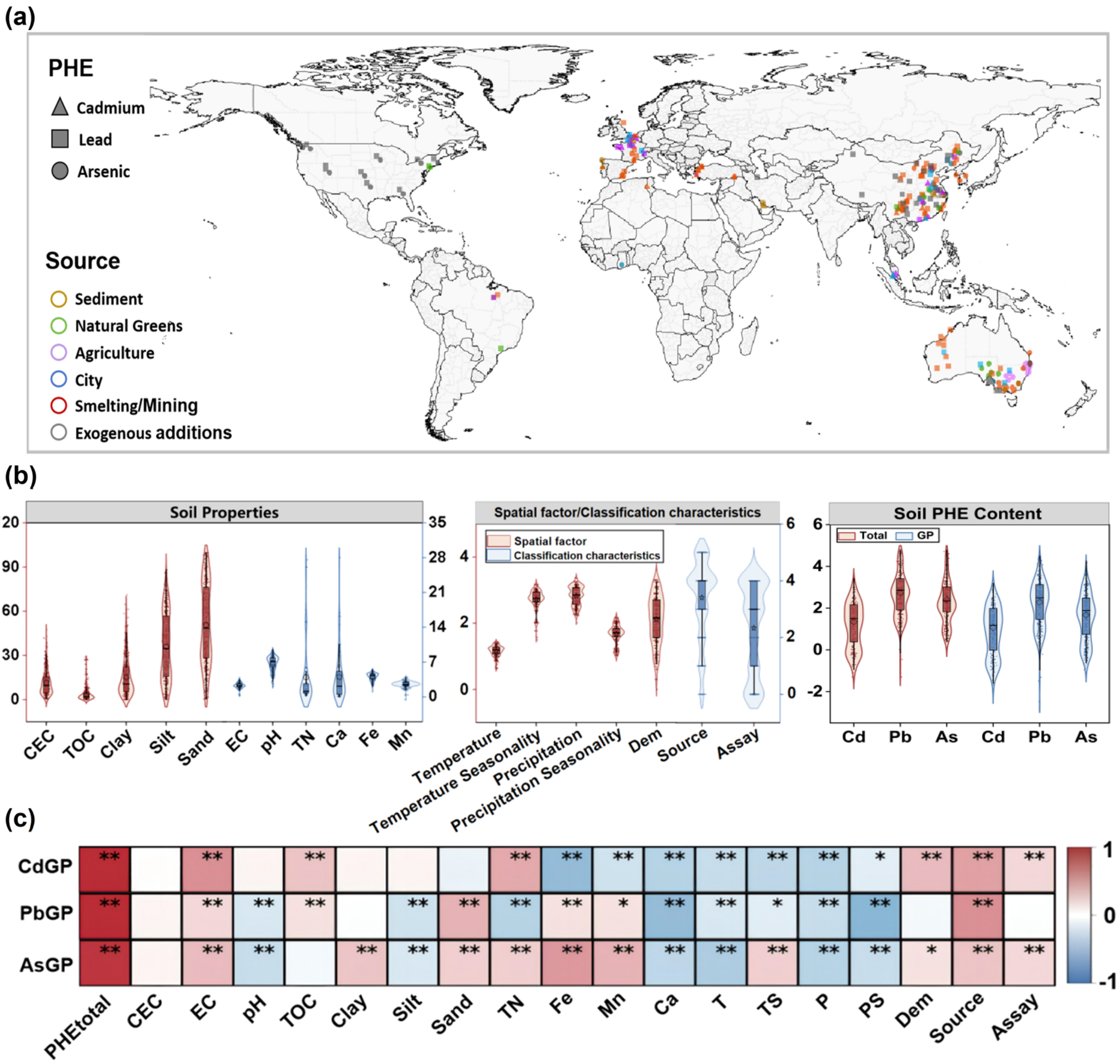


Figure 2. Spatial and statistical analysis of environmental data: (a) spatial distribution of modeling data; (b) violin plot of data distribution; (c) heatmaps of correlations of target variables (significant correlation marker "*": $p < 0.05$; "**": $p < 0.01$).

2. MATERIALS AND METHODS

2.1. Data Collection and Preprocessing

In our research, we merged data sets from China National Knowledge Infrastructure (CNKI) and Web of Science with a focus on key terms including "soil heavy metals," "Cd," "Pb," "As," "bioavailability," and "bioaccessibility" to craft an extensive literature database on soil heavy metal bioaccessibility. This compilation was enhanced by integrating diverse soil properties, environmental variables, and a range of in vitro simulation digestion methods to explore the connection between the bioaccessibility of PHE and soil attributes. Our data set encompasses 54 peer-reviewed papers from 2005 to 2023, offering predictions on the gastric bioaccessibility of PHE in soil. Additionally, we aggregated pollution concentration data from 151 studies covering all 31 Chinese provinces, amassing a total of 3068 data points. During the preprocessing phase, we addressed missing values, encoded categorical features, and identified and removed outliers. This process refined our data set into three targeted subsets: Cd, Pb, and As, laying the groundwork for model development, performance evaluation, and comprehensive analysis. For a more detailed methodology, description of the data set, and the analysis process, please refer to Appendix text S1.

2.2. Model Development and Interpretation

In our study, we crafted a comprehensive model that melded a variety of machine learning algorithms. These algorithms include linear models such as ridge regression (RR), decision tree methods such as RF, integration techniques, especially XGBoost (XGB) and CatBoost (CB), and multilayer perceptron (MLP) models. Intriguingly, the core algorithms of XGB and CB equipped them to adequately manage data sets with missing values.[31] Building on this, we implemented a fusion model anchored in the boosting ensemble learning strategy. Here, RR, RF, XGB, and CB served as the foundational learners, augmented by the gradient boosting algorithm.[32] For a granular understanding of this process, please refer to Appendix text S2. We then employed the finalized ensemble model on a test set, gauging its performance using the average coefficient of determination (R^2) and the root-mean-square error (RMSE) metrics across a spectrum of one hundred distinct random states. A 5-fold cross-validation was further used to evaluate its generalization ability (Figure 1).

In the field of machine learning, interpretations remain crucial. We not only utilized the built-in explanatory module of the gradient boosting integrated learning model to directly articulate the importance of features, thereby revealing their effects on the prediction results but also employed SHAP-based analysis of the importance of model features. The essence of the built-in explanation module is to assess the importance of features through metrics such as

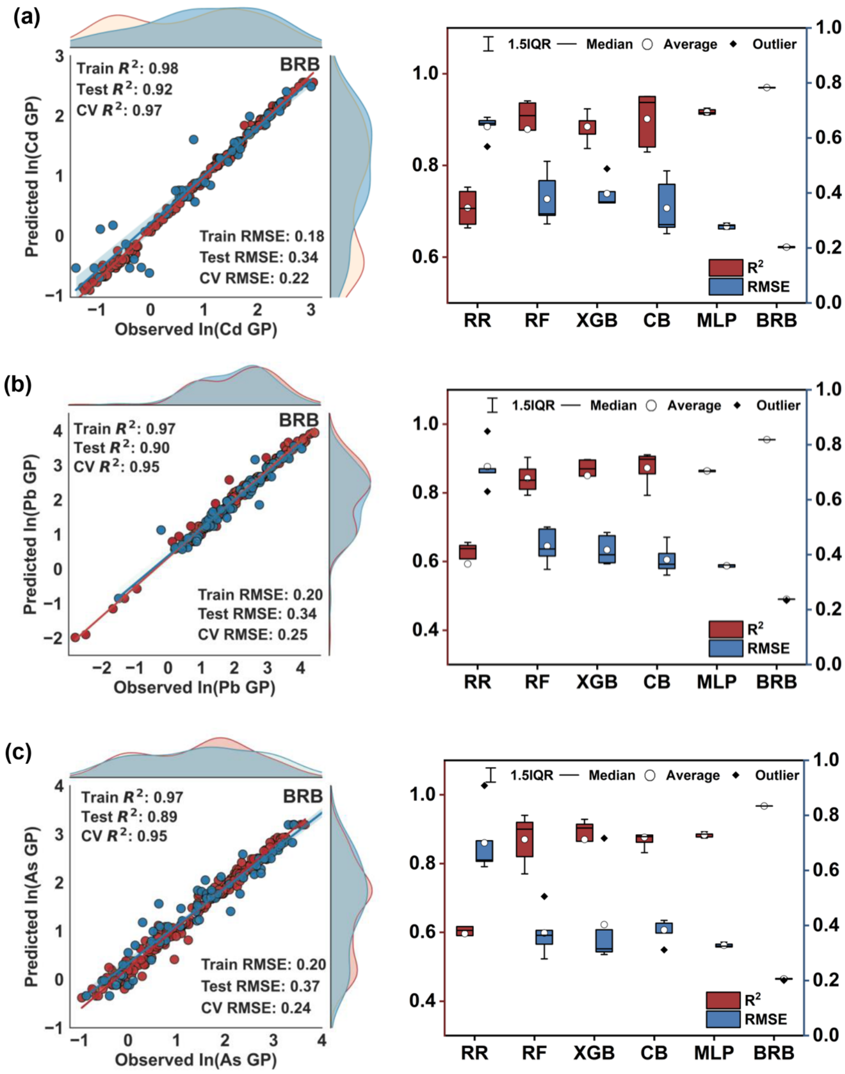


Figure 3. Model performance evaluation for (a) Cd; (b) Pb; (c) As. (red: training set; blue solid line: test set; light blue shaded padding: 95% confidence intervals for test set fit lines).

the frequency of tree node splits and subsequent gains. Although this approach is known for its simplicity, clarity, and efficiency, it mainly reveals the relative importance of features and neglects the interactions between them.[33] To bridge this knowledge gap, a quantitative analysis based on SHAP values was employed to study the cumulative impact of multiple features on the output.[34] A more in-depth exposition can be found in Appendix text S3. Furthermore, all the preprocessing, development, and interpretation of the models mentioned above were conducted using Python 3.7 through the Anaconda distribution, specifically leveraging the scikit-learn package (version 0.24.1).

2.3. Application of the Model

Leveraging our refined machine learning model, we projected the bioaccessibility of PHE across various Chinese regions and sites. Using the inverse distance weighting (IDW) technique in ArcMap, we transformed these projections into intuitive heatmaps, highlighting the spatial distribution of PHE bioaccessibility. By aligning these visualizations with the Chinese standards for soil pollution in construction areas, we established the SEC for representative sites (Tables S2−S3).[35] Evaluating these standards against concentration data from our data set offered insights into the model's accuracy and reliability. All analyses were performed in ArcMap 10.8.

3. RESULTS AND DISCUSSION

3.1. Descriptive Statistics for Modeled Data Sets

To gain a preliminary understanding of the original data set, we performed descriptive statistics analysis on all numerical features. Figure 2a illustrates the spatial distribution of all sampling points. Moreover, Figure 2b and Tables S4−5 depict the distribution patterns and approximate ranges of soil properties and spatial factors, respectively. The data set reveals significant variations in soil characteristics and spatial factors, covering most soil types in China. This confirms that our data set, gathered from the literature, is a representative sample for researching the regional bioaccessibility of Cd, Pb, and As in China's soils.

The bioaccessibility of PHE in the gastric phase, including $Cd_{gastric}$, $As_{gastric}$, and $Pb_{gastric}$, were chosen as the targets for our predictions. Their approximate ranges and distributions are presented in Table S6. Specifically, $Cd_{gastric}$ ranges from 0.01 to 1093.6 mg/kg with an average of 61.3 mg/kg, with magnitudes from 10^{-2} to $10.^{3}$ $As_{gastric}$ varies from 0.11 to 14775.95 mg/kg, averaging 313.8 mg/kg, and ranges from 10^{-1} to $10,^{4}$ showing a broader variance than $Cd_{gastric}$. $Pb_{gastric}$ has the widest range among the three, from 0.0015 to 26347.4 mg/kg, with

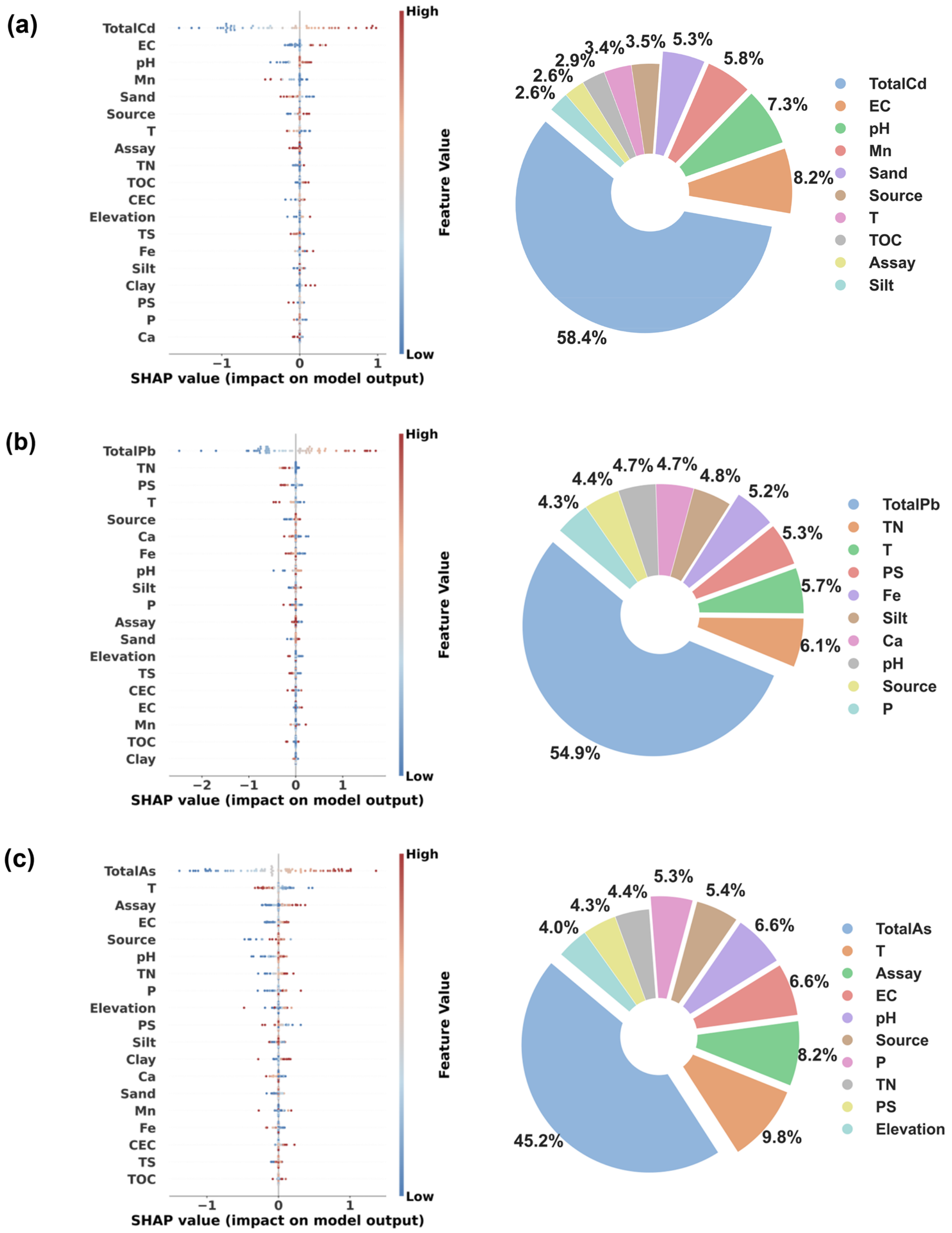


Figure 4. Shape waterfall and relative importance plots for the MLP model are presented for (a) Cd; (b) Pb; (c) As. (The color in the shape waterfall chart indicates the magnitude of the feature values, while the horizontal axis represents their contribution to the prediction (output). The relative importance chart depicts the marginal contribution of each feature to the output.).

magnitudes from 10^{-4} to 10^{5} and an average of 1186.3 mg/kg. Overall, the three prediction targets exhibit skewed distributions, with that of $\mathrm{Pb}_{\mathrm{gastric}}$ being the most pronounced. To address this skewness and enhance model performance, we applied a logarithmic transformation to our prediction targets.

Furthermore, the performance of the imputed numerical features after the KNN imputation was employed is depicted in Figure S1. A comparison between the distributions of the imputed and original data sets reveals a close resemblance, affirming the reliability of our imputation process. We further

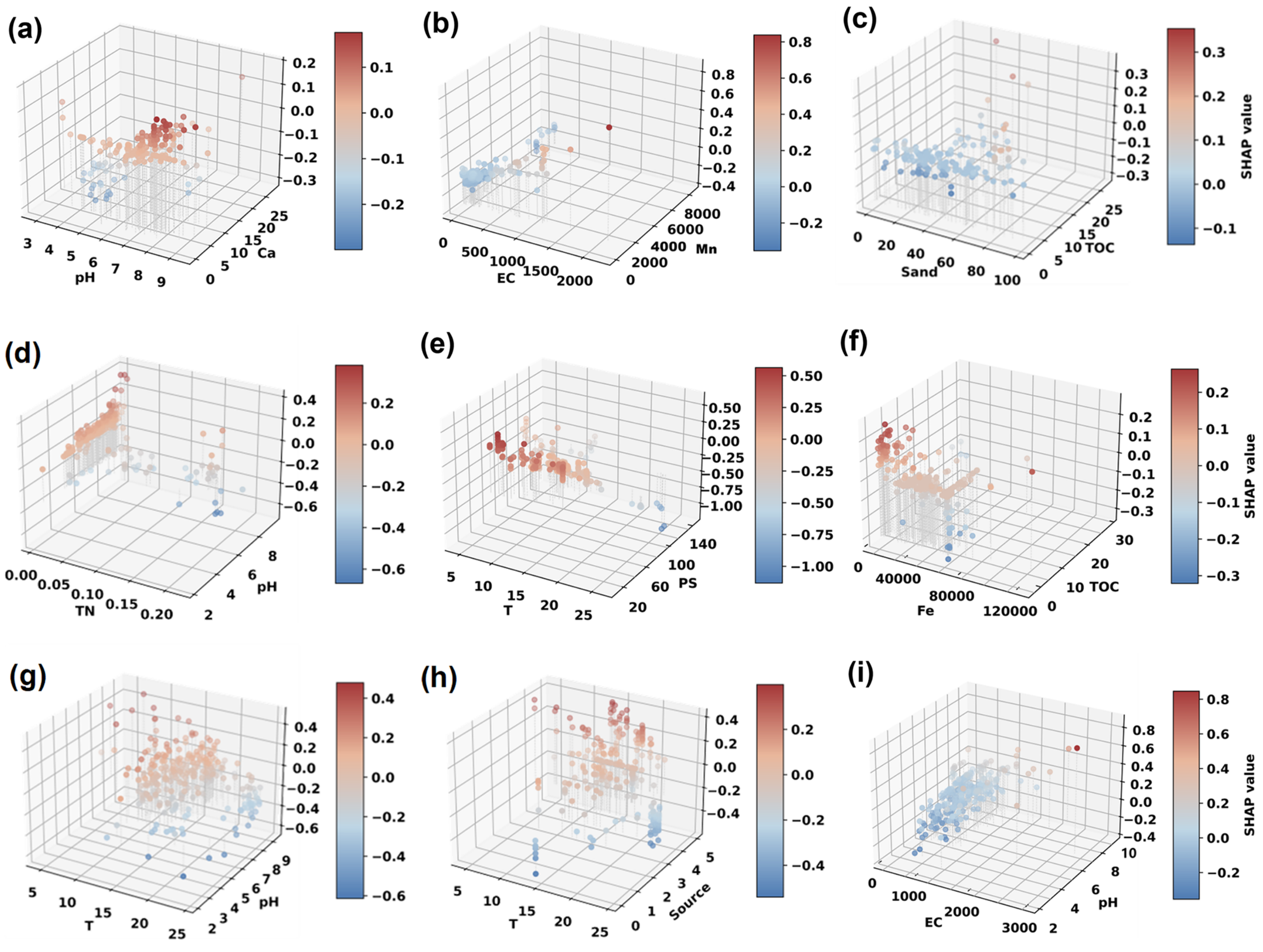


Figure 5. SHAP qualitative analysis. Red dots indicate that the combined parameter's total SHAP values are positive, while blue dots signify that the overall effects are negative. Points with different colors on the *z*-axis (SHAP values) represent the interaction of other factors. (a) The total SHAP value of pH and Ca. (b) The total SHAP value of EC and Mn. (c) The total SHAP value of sand and TOC. (d) The total SHAP value of TN and pH. (e) The total SHAP value of T and PS. (f) The total SHAP value of Fe and TOC. (g) The total SHAP value of T and pH. (h) The total SHAP value of TOC and source. (i) The total SHAP value of EC and pH.

explored the relationships between all feature variables and the prediction target within the imputed data set using Pearson's rank correlation analysis.[36] As illustrated in Figure 2c and Figure S2, there is a significant correlation between the predicted PHE bioaccessibility and soil properties. Additionally, climate and elevation data show varying degrees of correlation with the bioaccessibility of different PHE, suggesting their potential influence on the mobility and accumulation of PHE in soil.[36] Figure 2 also reveals a pronounced correlation between categorical features (source and assay) and the bioaccessibility of Cd and As, emphasizing the importance of considering these categorical features. While these features do exhibit significant correlations with PHE bioaccessibility, the majority of the correlation coefficients are relatively low. This indicates weak multicollinearity among them. Consequently, simple linear models may be insufficient to capture the complex relationships between these characteristics.[37]

3.2. Evaluation of the Performance of Models

After completing data imputation, we employed five machine learning algorithms, specifically RR, RF, XGB, CB, and MLP, utilizing 19 input features to predict the bioaccessibility of soil PHE in the three subdata sets for Cd, Pb, and As. Figure 3a depicts the predictive performance of different models after 5-fold cross-validation using R^2 and RMSE. By comparing the scores of the linear regression (RR), bagging (RF), boosting (XGB, CB), and artificial neural network (MLP) algorithms, we found that although the RR model outperformed the other models for all three subdata sets, there were still significant differences compared to the integrated and artificial neural network algorithms.[38] This might be attributed to the fact that linear models might not capture nonlinear relationships as effectively as other methods when the number of input features is high and their interrelationships become complex.[38] The performance of the MLP model was similar to that of the tree-based model for each of the subdata sets, while the integrated tree-based bagging and boosting algorithms performed slightly differently on the three subdata sets. Although both the boosting and bagging algorithms use a tree structure, there are significant differences in the way they are processed. The boosting algorithm focuses on hard-to-classify samples in the unbalanced data set to reduce the loss function and to enhance the weights of positively classified samples. The bagging algorithm, on the other hand, trains each tree independently through random sampling.[39] This could explain the observed scoring trends in the different models in our study.

Moreover, to leverage the strengths and compensate for the weaknesses of multiple models, we integrated the aforementioned five models using three ensemble strategies: voting, stacking, and boosting. As shown in Figure S3 and Figure 3, the boosting strategy outperformed the other two strategies and even surpassed the performance of individual models. Evaluating the cross-validation performance of the bioaccessibility prediction models (BRBs) for the three subdata sets allowed us to gauge their capacity to handle unknown data and generalize effectively. As illustrated in Figure 3, the BRB models for the three subdata sets produced R^2 scores of 0.97 (Cd), 0.95 (Pb), and 0.95 (As) after 100 random simulations of cross-validation. These scores closely align with the test set performance, highlighting the high predictive accuracy and generalization capability of the BRB models constructed for Cd, Pb, and As. This emphasizes the advantages of adopting an ensemble approach. Despite significant data volume differences across the three subdata sets, their model performances remained consistent, indicating our model's ability to capture inherent relationships effectively, even within limited data sets. Additionally, the optimal BRB models for these subdata sets achieved relatively low RMSE values both on the test set and during cross-validation, specifically Cd: 0.22, Pb: 0.25, and As: 0.25. The low variability in predicted values further underscores the model's superior performance in predicting the bioaccessibility of PHE in site soils.

3.3. Interpretation of the Fusion Model

In the PHE bioaccessibility models of the three subdata sets, we primarily analyzed the MLP model with the highest contribution rate in the integrated model through weight-based model-intrinsic methods and kernel SHAP calculation methods (Figure S4). Additionally, using a quantitative analysis, we further elucidated the features that are particularly important for predicting soil PHE bioaccessibility (Figure 4).

The bioaccessibility of potentially PHE in the gastric phase, including Cdgastric, Asgastric, and Pbgastric, were chosen as the targets for our predictions. Notably, across different interpretation strategies, the PHEtotal in the soil consistently played a central role in predicting PHE bioavailability. Moreover, our findings regarding the impact of parameters such as EC, pH, particle size distribution, CEC, and TOC on PHE bioaccessibility are in line with the literature, and these parameters have been widely acknowledged to be important among the academic community.[40,41] However, the kernel SHAP analysis also revealed that some spatial features and encoded categorical features were important (with a relative feature importance greater than 5%) in predicting Pb and As in soil. Additionally, based on the distribution of feature descriptors in the SHAP graph, we observed that the soil PHEtotal might exhibit a nonlinear positive correlation with bioaccessibility. Furthermore, the high feature values of EC and pH (marked in red) were primarily concentrated on the right side of the graph, suggesting a possible positive correlation with PHE bioaccessibility. Conversely, features such as Mn, sand, TN, and Fe had more high-value points on the left, emphasizing their potential negative correlation with PHE bioaccessibility (Figure 4). A detailed quantitative analysis of these key features is undertaken in subsequent sections.

As shown in Figure 5a, under acidic conditions, the stability of complexes formed by organic matter and hydroxyl ions with Cd increases as soil pH rises. At the same time, at high concentrations, protons compete with soil anion exchange sites, leading to the release of more Cd. Furthermore, Ca ions compete with these sites, enhancing the bioaccessibility of Cd in the soil.[42] However, as soil pH increases to alkaline levels, the number of insoluble compounds or precipitates formed by Cd with soil anions increases, leading to a decrease in its bioaccessibility.[43] Figure 5b indicates that with an increase in soil EC, soil adsorption of Cd is weakened.[44] However, although high ion concentrations potentially promote the transfer of Cd from the soil's solid phase to the aqueous phase, Mn oxides and hydroxides in the soil might form stable complexes with Cd, slightly limiting its mobility and bioaccessibility in the aqueous phase.[45,46] Additionally, Figure 5g reveals that the bioaccessibility of As in soil is highest under neutral pH and high EC conditions, confirming the aforementioned analysis. Furthermore, soil texture and elemental content can affect the bioaccessibility of PHE in soil. As depicted in Figure 5c, due to its lower surface area, the sand fraction has fewer adsorption sites for Cd, causing Cd to be released into the soil aqueous phase. The low water retention of sand expedites the leaching of Cd, reducing its bioaccessibility.[47] However, an increase in TOC in sandy soils enhances Cd adsorption, inhibiting the leaching of Cd from the soil.[48] Figure 5d illustrates the trend in soil Pb bioaccessibility with changes in TN. At a soil pH of 2−6, due to the protonation of nitrogen organic functional groups, the complexes formed by Pb^{2+} might be more stable. However, as the soil pH rises from 6 to 10, the number of negative charges increases due to functional group deprotonation, which weakens complex formation with Pb, resulting in increased soil Pb bioaccessibility at the same TN content. This further indicates that the main mechanism by which TN affects the bioaccessibility of Pb in the soil is the chelation effect of TN.[49] Figure 5f shows that an increase in Fe content results in the enhanced adsorption of Pb, reducing its bioaccessibility. However, when the soil TOC content (greater than 15%) is high enough to form complexes with Fe, TOC inhibits the oxidation of Fe, reducing Pb adsorption and thereby increasing its bioaccessibility.[50] It is noteworthy, as depicted in Figure 5g, that in the prediction of soil As bioaccessibility, there is an interaction between spatial variables (T) and soil properties (such as pH). Specifically, when the temperature is below 15 degrees, increasing pH has a significant impact on the SHAP value. Figure 5e further demonstrates that in cold and arid areas, the bioaccessibility of Pb is higher. This underscores the importance of considering spatial heterogeneity and its interaction with environmental factors when predicting PHE bioaccessibility. Moreover, in terms of categorical features, Figure 5h reveals that the bioaccessibility of As in natural lands with vegetation cover is lower, possibly because arsenic primarily exists in natural environments as stable compounds that do not absorb easily. In contrast, mining sites, due to long-term excavation and smelting activities, show an increase in the proportion of bioaccessibility soil As.

3.4. General Status of Heavy Metal Pollution in China

The descriptive statistics for PHE concentrations in China are presented in Table S7−8. The average concentrations of Cd, Pb, and As are 17.5 mg/kg, 451.2 mg/kg, and 175.5 mg/kg,

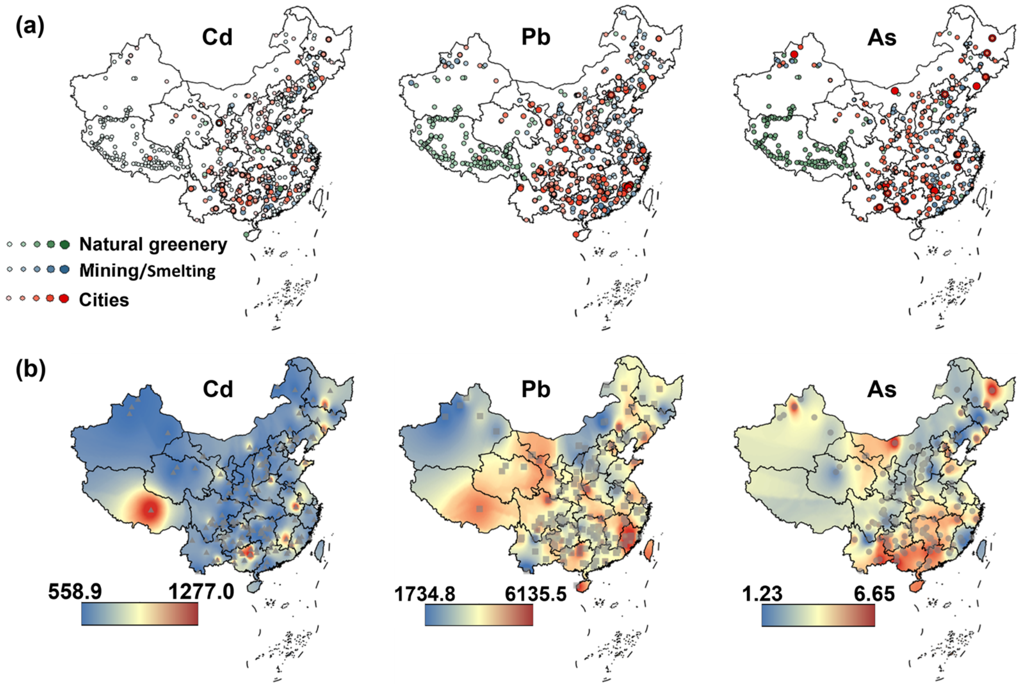


Figure 6. (a) Distribution of PHE contamination across various land use types in China. The concentration levels for Cd and As ranged from 0 to 1, 1−10, 10−100, 100−1000, and 1000−10000 mg/kg. For Pb, the levels range from 0 to 10, 10−100, 100−1000, 1000−10000, to 10000−100000 mg/kg. (b) Derivation of soil environmental standards (mg/kg) for Mining/Smelting sites in China based on the BRB model.

respectively. When compared with the current soil environmental standards in China, which set values for Cd, Pb, and As at 65, 800, and 60 mg/kg respectively, exceedance rates vary across land use types. Specifically, for Cd, the rates are 0.9% in natural greens (Ng), 6.2% in city industry (Ci), and 7.2% in mining and smelting (Ms). For Pb, the rates are 0% in Ng, 2.4% in Ci, and 19.1% in Ms. As for As, rates of exceedance are recorded at 1.1% in Ng, 10.2% in Ci, and notably, 34.2% in Ms[7]. This indicates that areas used for mining/smelting have the most significant contamination, followed by urban commercial areas and natural green spaces. Furthermore, natural green spaces have a lower coefficient of variation (CV), which suggests that the elemental concentrations are primarily influenced by the inherent soil background and natural processes.[51,52] In contrast, in mining/smelting and urban commercial areas in China, where the CV for PHE concentrations exceeds 2, there is pronounced spatial variation in PHE contamination. This highlights the presence of point-source pollution, with human activities significantly contributing to pollution in these areas.[53]

Figure 6a depicts the distribution patterns of Cd, Pb, and As across China. Notably, significant PHE pollution is concentrated in the southeastern region of China. Specifically, key PHE contamination clusters can be identified at the confluence of Guizhou and Yunnan, the Hunan-Guangdong border, and Guangxi's Heyuan area. These hotspots likely result from historical smelting, a prevalence of ore deposits, and unchecked industrial waste discharges.[54] Furthermore, parts of Fujian show notable lead pollution, primarily from smelting (Figure 6a). This heightened presence of lead is potentially due to past industrial and mining activities.[55] Farther northward in the southeast, isolated mining and smelting sites emerge as pollution focal points. It is important to highlight that even in the northwestern areas, bordered by the Heihe-Tengchong line, sporadic PHE pollution exists. For example, Gansu's lead contamination is linked to its lead−zinc smelting plants (Figure 6a). Scattered arsenic pollution in the northwestern and northeastern provinces might be linked to industries such as coal mining (Figure 6a).[56] Given these insights, it is imperative to devise targeted soil management strategies for mining and smelting areas, ensuring robust environmental and public safety.

3.5. Determination of Soil Environmental Criteria for Chinese Sites Based on the BRB Model

To evaluate the performance of our model in site SEC derivation, we utilized the previously trained BRB model to forecast the bioaccessibility distribution of Cd, Pb, and As in the three subdata sets in China. As shown in Figure S6, the analysis of 3002 samples from across the country indicates that the bioaccessibility of cadmium is significantly higher than that of lead and arsenic. Additionally, the bioaccessibility of lead and arsenic is notably increased in northwestern parts of China, at the border between Shandong and Henan, and along the southeastern coast. The elevated bioaccessibility of soil Cd might be attributed to the BRB model's inability to predict PHE bioaccessibility at extremely low concentrations due to a lack of training data. As a result, biases cause the predicted PHE bioaccessibility to exceed 100%. Considering the negligible health impact of extremely low soil PHE concentrations on the public, we set the bioaccessibility to 100% when deriving the soil environmental standards for these areas. Although our model reveals regional disparities in the bioaccessibility of Cd, Pb, and As in soils, its conclusions are limited by a small data set. Future research, enriched by additional data, will improve the model's accuracy and broaden its applicability.

Furthermore, based on inverse distance weighting interpolation, we mapped the bioaccessibility thermal distribution in various locations across China (Figure S6). To enhance the accuracy of health risk assessments for PHE in site soil, we applied formulas from Table S9, transforming bioaccessibility into bioavailability. Moreover, drawing upon the "Technical Guidelines for Risk Assessment of Soil Contamination of Construction Land" currently in force in China, we developed SEC tailored to various PHE found in soils of industrial sites.[35] The criteria for the three PHE are as follows: Cd (558.99−1277.04 mg/kg), Pb (1703.37−6281.43 mg/kg), and As (1.23−6.65 mg/kg), as shown in Figure 6b. These criteria exhibit clear regional patterns. Our model captures dynamic differences in soil PHE bioaccessibility across different regions in China, enabling derivations of soil environmental criteria across continental scales.

In addition, we conducted an in-depth analysis of the samples from the three subdata sets of metallurgical site types. As demonstrated in Figure S7, compared to the criteria recommended by the current Chinese guidelines (Table S10), the number of samples that exceeded the standards for Cd, Pb, and As decreased from 5, 58, and 14 to 1, 24, and 7, respectively. This significant reduction indicates our method's precision in pinpointing areas that are at genuine risk. It is estimated that during China's "13th Five-Year Plan," over 3 million acres of metallurgical land was set aside for restoration, representing a market potential of over 300 billion Chinese yuan.[57] Implementing our model could offer governmental departments and decision-making bodies a more precise screening of soil contamination, reducing unnecessary monitoring costs and soil remediation investments. In comparison to traditional assessment methods, our model provides a finer spatial resolution and higher data sensitivity, enabling us to better identify and locate potential pollution hotspots.

4. CONCLUSION

In this study, we analyzed 937 data points from 53 studies to showcase the potential of machine learning models for predicting PHE bioaccessibility in soil. The boosting method excelled, achieving an average R^2 of 0.95 and RMSE of 0.25 across three data sets. Using Kernel SHAP values, we identified how soil properties and spatial variables impact PHE bioaccessibility. The spatial distribution of PHE bioaccessibility throughout China was illustrated by the constructed machine learning model associated with Inverse Distance Weighting, uncovering that Cd bioaccessibility surpasses that of Pb and As, particularly in the northwestern and southeastern coastal areas. This insight laid the groundwork for formulating soil environmental standards for smelting/mining sites across China, significantly reducing the prevalence of sites exceeding the standards. However, limitations include reliance on lab data and insufficient low-concentration data. Incorporating deep feature analysis is crucial for model enhancement, with a focus on detailed classification of smelting/mining sites for a clearer understanding of industrial activity impacts. To improve, we advocate for expanding data sets and collection methods to comprehensively map PHE bioaccessibility in soil for better environmental health insights.

■ ASSOCIATED CONTENT

Data Availability Statement

The source codes employed, as well as a notebook to reproduce our results, and the instruction for using the package multiple-output model can be found on the GitHub: https://github.com/Kingsely-o/ML.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/envhealth.4c00035.

Detailed methodologies, analyses, and visualizations for studying soil PHE bioaccessibility in China; data collection, machine learning optimization, and risk assessment guidelines (PDF)

Modeling data (XLS)

Modeling application data (XLS)

■ AUTHOR INFORMATION

Corresponding Author

Yong Yuan − *Guangdong Key Laboratory of Environmental Catalysis and Health Risk Control, School of Environmental Science and Engineering, Institute of Environmental Health and Pollution Control, Guangdong University of Technology, Guangzhou 510006, China;* orcid.org/0000-0003-1513-9542; Email: yuanyong@soil.gd.cn

Authors

Kunting Xie − *Guangdong Key Laboratory of Environmental Catalysis and Health Risk Control, School of Environmental Science and Engineering, Institute of Environmental Health and Pollution Control, Guangdong University of Technology, Guangzhou 510006, China*

Jiajun Ou − *School of Automation, Guangdong University of Technology, Guangzhou 510006, China*

Minghao He − *Guangdong Key Laboratory of Environmental Catalysis and Health Risk Control, School of Environmental Science and Engineering, Institute of Environmental Health and Pollution Control, Guangdong University of Technology, Guangzhou 510006, China*

Weijie Peng − *Guangdong Key Laboratory of Environmental Catalysis and Health Risk Control, School of Environmental Science and Engineering, Institute of Environmental Health and Pollution Control, Guangdong University of Technology, Guangzhou 510006, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/envhealth.4c00035

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China (2019YFC1804604).

■ REFERENCES

(1) Shaheen, S. M.; El-Naggar, A.; Antoniadis, V.; Moghanm, F. S.; Zhang, Z.; Tsang, D. C. W.; Ok, Y. S.; Rinklebe, J. Release of toxic elements in fishpond sediments under dynamic redox conditions: Assessing the potential environmental risk for a safe management of fisheries systems and degraded waterlogged sediments. *J. Environ. Manage.* **2020**, *255*, 109778.

(2) Liu, X.; Chen, S.; Yan, X.; Liang, T.; Yang, X.; El-Naggar, A.; Liu, J.; Chen, H. Evaluation of potential ecological risks in potential toxic elements contaminated agricultural soils: Correlations between soil contamination and polymetallic mining activity. *J. Environ. Manage.* **2021**, *300*, 113679.

(3) Yang, X.; Li, J.; Liang, T.; Yan, X.; Zhong, L.; Shao, J.; El-Naggar, A.; Guan, C.-Y.; Liu, J.; Zhou, Y. A combined management scheme to

simultaneously mitigate As and Cd concentrations in rice cultivated in contaminated paddy soil. *J. Hazard. Mater.* **2021**, *416*, 125837.

(4) *National Soil Contamination Survey Report*; Ministry of Environmental Protection of the People's Republic of China, 2014.

(5) State Council of China Soil Pollution Prevention and Cleanup Action Plan. In *CNKI*, **2016**; p 8.

(6) Bone, J.; Head, M.; Jones, D. T.; Barraclough, D.; Archer, M.; Scheib, C.; Flight, D.; Eggleton, P.; Voulvoulis, N. From Chemical Risk Assessment to Environmental Quality Management: The Challenge for Soil Protection. *Environ. Sci. Technol.* **2011**, *45* (1), 104−110.

(7) *Soil Environmental Quality: Risk Control Standard for Soil Contamination of Development Land*; Ministry of Environment of the People's Republic of China, 2018.

(8) Antoniadis, V.; Shaheen, S. M.; Levizou, E.; Shahid, M.; Niazi, N. K.; Vithanage, M.; Ok, Y. S.; Bolan, N.; Rinklebe, J. A critical prospective analysis of the potential toxicity of trace element regulation limits in soils worldwide: Are they protective concerning health risk assessment?-A review. *Environ. Int.* **2019**, *127*, 819−847.

(9) İpek, M.; Ünlü, K. Development of human health risk-based Soil Quality Standards for Turkey: Conceptual framework. *Environ. Adv.* **2020**, *1*, 100004.

(10) Li, H.-B.; Li, M.-Y.; Zhao, D.; Li, J.; Li, S.-W.; Juhasz, A. L.; Basta, N. T.; Luo, Y.-M.; Ma, L. Q. Oral bioavailability of As, Pb, and Cd in contaminated soils, dust, and foods based on animal bioassays: a review. *Environ. Sci. Technol.* **2019**, *53* (18), 10545−10559.

(11) Yin, N.; Li, Y.; Cai, X.; Du, H.; Wang, P.; Han, Z.; Sun, G.; Cui, Y. The role of soil arsenic fractionation in the bioaccessibility, transformation, and fate of arsenic in the presence of human gut microbiota. *J. Hazard. Mater.* **2021**, *401*, 123366.

(12) Juhasz, A. L.; Smith, E.; Weber, J.; Rees, M.; Rofe, A.; Kuchel, T.; Sansom, L.; Naidu, R. Comparison of in vivo and in vitro methodologies for the assessment of arsenic bioavailability in contaminated soils. *Chemosphere.* **2007**, *69* (6), 961−966.

(13) Juhasz, A. L.; Weber, J.; Naidu, R.; Gancarz, D.; Rofe, A.; Todor, D.; Smith, E. Determination of cadmium relative bioavailability in contaminated soils and its prediction using in vitro methodologies. *Environ. Sci. Technol.* **2010**, *44* (13), 5240−5247.

(14) Smith, E.; Kempson, I. M.; Juhasz, A. L.; Weber, J.; Rofe, A.; Gancarz, D.; Naidu, R.; McLaren, R. G.; Gräfe, M. In vivo−in vitro and XANES spectroscopy assessments of lead bioavailability in contaminated periurban soils. *Environ. Sci. Technol.* **2011**, *45* (14), 6145−6152.

(15) Li, J.; Li, K.; Cave, M.; Li, H.-B.; Ma, L. Q. Lead bioaccessibility in 12 contaminated soils from China: Correlation to lead relative bioavailability and lead in different fractions. *J. Hazard. Mater.* **2015**, *295*, 55−62.

(16) Li, S.-W.; Sun, H.-J.; Li, H.-B.; Luo, J.; Ma, L. Q. Assessment of cadmium bioaccessibility to predict its bioavailability in contaminated soils. *Environ. Int.* **2016**, *94*, 600−606.

(17) Li, H.-B.; Li, M.-Y.; Zhao, D.; Li, J.; Li, S.-W.; Xiang, P.; Juhasz, A. L.; Ma, L. Q. Arsenic, lead, and cadmium bioaccessibility in contaminated soils: measurements and validations. *Crit. Rev. Environ. Sci. Technol.* **2020**, *50* (13), 1303−1338.

(18) Li, J.; Li, K.; Cui, X.-Y.; Basta, N. T.; Li, L.-P.; Li, H.-B.; Ma, L. In vitro bioaccessibility and in vivo relative bioavailability in 12 contaminated soils: Method comparison and method development. *Sci. Total Environ.* **2015**, *532*, 812−820.

(19) Monneron–Gyurits, M.; Soubrand, M.; Joussein, E.; Courtin-Nomade, A.; Jubany, I.; Casas, S.; Bahi, N.; Faz, A.; Gabarron, M.; Acosta, J. A.; Martinez-Martinez, S. Investigating the relationship between speciation and oral/lung bioaccessibility of a highly contaminated tailing: contribution in health risk assessment. *Environ. Sci. Pollut. Res.* **2020**, *27*, 40732−40748.

(20) Amnai, A.; Radola, D.; Choulet, F.; Buatier, M.; Gimbert, F. Impact of ancient iron smelting wastes on current soils: Legacy contamination, environmental availability and fractionation of metals. *Sci. Total Environ.* **2021**, *776*, 145929.

(21) Liu, Y.; Du, Q.; Wang, Q.; Yu, H.; Liu, J.; Tian, Y.; Chang, C.; Lei, J. Causal inference between bioavailability of heavy metals and environmental factors in a large-scale region. *Environ. Pollut.* **2017**, *226*, 370−378.

(22) Zhang, C.; Yu, Z.-g.; Zeng, G.-m.; Jiang, M.; Yang, Z.-z.; Cui, F.; Zhu, M.-y.; Shen, L.-q.; Hu, L. Effects of sediment geochemical properties on heavy metal bioavailability. *Environ. Int.* **2014**, *73*, 270−281.

(23) Juhasz, A. L.; Basta, N. T.; Smith, E. What is required for the validation of in vitro assays for predicting contaminant relative bioavailability? Considerations and criteria. *Environ. Pollut.* **2013**, *180*, 372−375.

(24) Yan, K.; Dong, Z.; Wijayawardena, M. A.; Liu, Y.; Li, Y.; Naidu, R. The source of lead determines the relationship between soil properties and lead bioaccessibility. *Environ. Pollut.* **2019**, *246*, 53−59.

(25) Pelfrene, A.; Waterlot, C.; Mazzuca, M.; Nisse, C.; Cuny, D.; Richard, A.; Denys, S.; Heyman, C.; Roussel, H.; Bidar, G.; Douay, F. Bioaccessibility of trace elements as affected by soil parameters in smelter-contaminated agricultural soils: a statistical modeling approach. *Environ. Pollut.* **2012**, *160*, 130−138.

(26) Yang, H.; Huang, K.; Zhang, K.; Weng, Q.; Zhang, H.; Wang, F. Predicting heavy metal adsorption on soil with machine learning and mapping global distribution of soil adsorption capacities. *Environ. Sci. Technol.* **2021**, *55* (20), 14316−14328.

(27) Wei, J.; Liu, S.; Li, Z.; Liu, C.; Qin, K.; Liu, X.; Pinker, R. T.; Dickerson, R. R.; Lin, J.; Boersma, K.; et al. Ground-Level NO_2 Surveillance from Space Across China for High Resolution Using Interpretable Spatiotemporally Weighted Artificial Intelligence. *Environ. Sci. Technol.* **2022**, *56* (14), 9988−9998.

(28) Liao, Z.; Lu, J.; Xie, K.; Wang, Y.; Yuan, Y. Prediction of Photochemical Properties of Dissolved Organic Matter Using Machine Learning. *Environ. Sci. Technol.* **2023**, *57* (46), 17971−17980.

(29) Xie, K.; Xie, N.; Liao, Z.; Luo, X.; Peng, W.; Yuan, Y. Bioaccessibility of arsenic, lead, and cadmium in contaminated mining/smelting soils: Assessment, modeling, and application for soil environment criteria derivation. *J. Hazard. Mater.* **2023**, *443*, 130321.

(30) Zhang, B.; Liu, L.; Huang, Z.; Hou, H.; Zhao, L.; Sun, Z. Application of stochastic model to assessment of heavy metal (loid) s source apportionment and bio-availability in rice fields of karst area. *Sci. Total Environ.* **2021**, *793*, 148614.

(31) Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K.; Mitchell, R.; Cano, I.; Zhou, T. Xgboost: extreme gradient boosting. *R package version 0.4−2* **2015**, *1* (4), 1−4.

(32) Bentéjac, C.; Csörgő, A.; Martínez-Muñoz, G. A comparative analysis of gradient boosting algorithms. *Artif. Intell.Rev.* **2021**, *54*, 1937−1967.

(33) Zhao, S.; Wang, M.; Ma, S.; Cui, Q. A feature selection method via relevant-redundant weight. *Expert Syst. Appl.* **2022**, *207*, 117923.

(34) He, B.; Zhu, X.; Cang, Z.; Liu, Y.; Lei, Y.; Chen, Z.; Wang, Y.; Zheng, Y.; Cang, D.; Zhang, L. Interpretation and Prediction of the CO2 Sequestration of Steel Slag by Machine Learning. *Environ. Sci. Technol.* **2023**, *57* (46), 17940−17949.

(35) *Technical Guidelines for Risk Assessment of Contaminated Sites (HJ25.3-2019)*; Ministry of Environment of the People's Republic of China, 2019.

(36) Chen, H.; Teng, Y.; Lu, S.; Wang, Y.; Wang, J. Contamination features and health risk of soil heavy metals in China. *Sci. Total Environ.* **2015**, *512*, 143−153.

(37) Farrar, D. E.; Glauber, R. R. Multicollinearity in regression analysis: the problem revisited. *Review of Economic Statistics* **1967**, *49*, 92−107.

(38) Hastie, T.; Tibshirani, R.; Friedman, J. H.; Friedman, J. H. *The elements of statistical learning: data mining, inference, and prediction*; Springer, 2009.

(39) Golden, C. E.; Rothrock, M. J., Jr; Mishra, A. Comparison between random forest and gradient boosting machine methods for predicting Listeria spp. prevalence in the environment of pastured poultry farms. *Food Res. Int.* **2019**, *122*, 47−55.

(40) Alloway, B. J. *Heavy metals in soils: trace metals and metalloids in soils and their bioavailability*; Springer Science & Business Media, 2012.

(41) Griggs, J. L.; Thomas, D. J.; Fry, R.; Bradham, K. D. Improving the predictive value of bioaccessibility assays and their use to provide mechanistic insights into bioavailability for toxic metals/metalloids—A research prospectus. *J. Toxicol. Environ. Health, Part B* **2021**, *24* (7), 307−324.

(42) Qin, F.; Shan, X.-q.; Wei, B. Effects of low-molecular-weight organic acids and residence time on desorption of Cu, Cd, and Pb from soils. *Chemosphere.* **2004**, *57* (4), 253−263.

(43) Shahid, M.; Dumat, C.; Khalid, S.; Niazi, N. K.; Antunes, P. M. Cadmium bioavailability, uptake, toxicity and detoxification in soil-plant system. *Rev. Environ. Contam. Toxicol.* **2016**, *241*, 73−137.

(44) Acosta, J.; Jansen, B.; Kalbitz, K.; Faz, A.; Martínez-Martínez, S. Salinity increases mobility of heavy metals in soils. *Chemosphere.* **2011**, *85* (8), 1318−1324.

(45) Wang, W.; Lu, T.; Liu, L.; Yang, X.; Sun, X.; Qiu, G.; Hua, D.; Zhou, D. Zeolite-supported manganese oxides decrease the Cd uptake of wheat plants in Cd-contaminated weakly alkaline arable soils. *J. Hazard. Mater.* **2021**, *419*, 126464.

(46) Zhang, Y.; Li, A.; Liu, L.; Duan, X.; Ge, W.; Liu, C.; Qiu, G. Enhanced remediation of cadmium-polluted soil and water using facilely prepared MnO2-coated rice husk biomass. *Chem. Eng. J.* **2023**, *457*, 141311.

(47) Huang, B.; Yuan, Z.; Li, D.; Zheng, M.; Nie, X.; Liao, Y. Effects of soil particle size on the adsorption, distribution, and migration behaviors of heavy metal (loid) s in soil: A review. *Environ. Sci. Processes Impacts.* **2020**, *22* (8), 1596−1615.

(48) Xu, Z.; Lu, Z.; Zhang, L.; Fan, H.; Wang, Y.; Li, J.; Lin, Y.; Liu, H.; Guo, S.; Xu, M.; Wang, J. Red mud based passivator reduced Cd accumulation in edible amaranth by influencing root organic matter metabolism and soil aggregate distribution. *Environ. Pollut.* **2021**, *275*, 116543.

(49) Fan, X.; Wang, X.; Cai, Y.; Xie, H.; Han, S.; Hao, C. Functionalized cotton charcoal/chitosan biomass-based hydrogel for capturing Pb2+, Cu2+ and MB. *J. Hazard. Mater.* **2022**, *423*, 127191.

(50) Zhao, Q.; Poulson, S. R.; Obrist, D.; Sumaila, S.; Dynes, J. J.; McBeth, J. M.; Yang, Y. Iron-bound organic carbon in forest soils: quantification and characterization. *Biogeosciences* **2016**, *13* (16), 4777−4788.

(51) Wang, H.; Yilihamu, Q.; Yuan, M.; Bai, H.; Xu, H.; Wu, J. Prediction models of soil heavy metal (loid) s concentration for agricultural land in Dongli: A comparison of regression and random forest. *Ecol. Indic.* **2020**, *119*, 106801.

(52) Azizi, K.; Ayoubi, S.; Nabiollahi, K.; Garosi, Y.; Gislum, R. Predicting heavy metal contents by applying machine learning approaches and environmental covariates in west of Iran. *J. Geochem. Explor.* **2022**, *233*, 106921.

(53) Hengl, T.; Nussbaum, M.; Wright, M. N.; Heuvelink, G. B.; Gräler, B. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ. Comput. Sci.* **2018**, *6*, No. e5518.

(54) Li, T.; Yu, X.; Li, M.; Rong, L.; Xiao, X.; Zou, X. Ecological insight into antibiotic resistome of ion-adsorption rare earth mining soils from south China by metagenomic analysis. *Sci. Total Environ.* **2023**, *872*, 162265.

(55) Xiaoniu, X. Geochronology, geochemistry and geological characteristics of granites from the Meixian zinc-lead polymetallic deposit in central Fujian Province. *Earth Sci. Front.* **2020**, *27* (4), 158.

(56) Zhou, S.; Wei, W.; Chen, L.; Zhang, Z.; Liu, Z.; Wang, Y.; Kong, J.; Li, J. Impact of a coal-fired power plant shutdown campaign on heavy metal emissions in China. *Environ. Sci. Technol.* **2019**, *53* (23), 14063−14069.

(57) Intelligence Research Group. *Market Demand Forecast and Investment Future Development Trend Report of China's Site Remediation Industry from 2019 to 2025*; Intelligence Research Group, 2019. https://www.chyxx.com/research/201907/764124.html (accessed 2023-08-17).