Journal of Hazardous Materials 463 (2024) 132910

Contents lists available at ScienceDirect

Journal of Hazardous Materials

journal homepage: www.elsevier.com/locate/jhazmat

Machine learning-based prediction of cadmium bioaccumulation capacity and associated analysis of driving factors in tobacco grown in Zunyi City, China

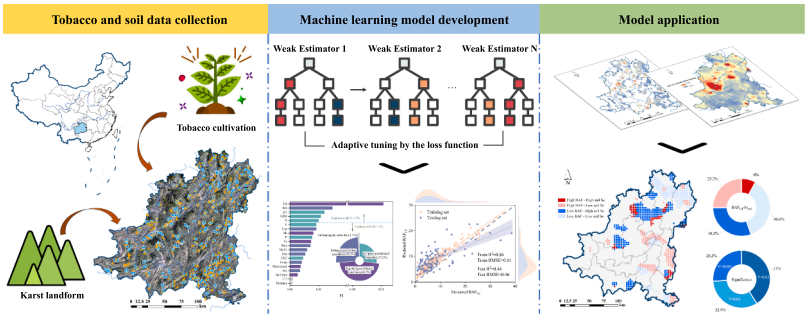
Zilun Gou [a,b,c], Chengshuai Liu [a], Meng Qi [a,c], Wenhao Zhao [b], Yi Sun [b], Yajing Qu [b], Jin Ma [b,*]

[a] State Key Laboratory of Environmental Geochemistry, Institute of Geochemistry, Chinese Academy of Sciences, Guiyang 550081, China
[b] State Key Laboratory of Environmental Criteria and Risk Assessment, Chinese Research Academy of Environmental Sciences, Beijing 100012, China
[c] University of Chinese Academy of Sciences, Beijing 100049, China

HIGHLIGHTS

- The spatial distribution of the BAF_{Cd} is clearly uneven among the study areas.
- The effect of anthropogenic activities on BAF_{Cd} is relatively insignificant.
- Elevated soil Se contents have an undeniable effect on the BAF_{Cd} in tobacco.
- LISA analysis is performed to identify the areas suitable for tobacco cultivation.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Wagner L. Araújo

Keywords:
High-geochemical background
Tobacco-growing soil
Machine learning
Local indicators of spatial association analysis

ABSTRACT

Tobacco grown in areas with high-geochemical backgrounds exhibits considerably different cadmium (Cd) bioaccumulation abilities due to regional disparities and environmental changes. However, the impact of key factors on the Cd bioaccumulation ability of tobacco grown in the karst regions with high selenium (Se) geochemical backgrounds is unclear. Herein, 365 paired rhizospheric soil-grown tobacco samples and 321 topsoil samples were collected from typical karst tobacco-growing soil in southwestern China and analyzed for Cd and Se. XGBoost was used to predict and evaluate the Cd bioaccumulation ability of tobacco and potential influencing factors. Results showed that regional geochemical characteristics, such as soil Cd and Se contents, soil type, and lithology, have the highest influence on the Cd bioaccumulation ability of tobacco, accounting for 46.5% of the overall variation. Moreover, soil Se contents in high-geochemical background areas considerably affect Cd bioaccumulation in tobacco, with a threshold for the mutual suppression effects of Cd and Se at a soil Se content of 0.8 mg/kg. According to the results of bivariate local indicators of spatial association analysis, tobacco cultivated in the central, northeast, and southeast regions of Zunyi City carries a lower risk of soil Cd contamination. This study provides new insights for managing tobacco cultivation in karst regions.

* Corresponding author.
E-mail address: majin@craes.org.cn (J. Ma).

https://doi.org/10.1016/j.jhazmat.2023.132910
Received 7 August 2023; Received in revised form 17 October 2023; Accepted 30 October 2023
Available online 3 November 2023

1. Introduction

The karst region in southwestern China is the primary area for the development of carbonate rocks [14]. The enormous volume effect during intense weathering and pedogenesis has resulted in a naturally high background of soil trace elements, particularly Cd, with a content as high as 1.21 mg/kg [37], which has severely impacted agricultural productivity in the karst region [15]. Tobacco is the main cash crop in the high-geochemical background areas of southwestern China, with a cultivation area of 38,000 ha and a yield of 66,000 t. Unfortunately, tobacco exhibits a strong tendency to bioaccumulate Cd, particularly in leaves [20,36]. As a result, smokers would be exposed to Cd during the smoking process, which has been proved to pose a severe threat to human health [25,27].

The presence of Cd in tobacco is predominantly attributed to its natural ability to bioaccumulate contaminants from soils [18]. Extensive research has been conducted on the soil-tobacco system based on this understanding. The observations from laboratory-scale cultivation experiments indicate that soil characteristics, such as soil pH and soil organic matter (SOM) content, play a crucial role in regulating Cd bioavailability and the bioaccumulation ability of Cd greatly varies under different pH conditions [16]. Moreover, essential elements, including nitrogen (N), phosphorus (P), and potassium (K), can modulate plant growth and gene expression, influencing the Cd tolerance of tobacco [27,38]. The results of a regional-scale survey indicate that Cd contents and bioavailabilities are considerably different in different lithologies [34]. Particularly, the development of carbonate rocks increases soil pH, which decreases Cd bioavailability [23]. Furthermore, the interactions among trace elements generally have a significant impact on the accumulation of Cd in tobacco. Among all trace elements, Se has attracted widespread attention because of its antagonism and synergism with Cd [5]. Guizhou Province is a region characterized by the highest soil Se content in China [19]. However, few studies have investigated the impact of various factors on the Cd bioaccumulation ability of tobacco in Se-rich areas because of limitations in conventional methods.

Conventional methods generally face difficulties in accurately deciphering geochemical processes. On the one hand, laboratory-scale experiments focusing on controlling environmental conditions to investigate the impact of a single factor on tobacco's ability to accumulate Cd may not fully reflect the true geochemical characteristics of the region. On the other hand, extensive soil sample collection is time-consuming and labor-intensive at the regional scale [39]. Furthermore, the use of traditional geostatistical analyses, such as interpolation techniques based on spatial autocorrelation, is highly reliant on observations near the prediction location and affected by the number of sampling points and soil heterogeneity, making it challenging to accurately assess tobacco's ability to bioaccumulate Cd [1,8]. Although attempts have been made in recent years to establish predictive models for soil properties and plant Cd content using statistical models such as multiple linear regression, such models only provide a simple linear summation of the various coupled forms between environmental covariates and tobacco's ability to bioaccumulate Cd [11,32]. Indeed, the relationship between driving factors and Cd accumulation in plants is often intricate and nonlinear [23]. Furthermore, nonnumeric variables such as soil type and parent rock, which may significantly affect Cd bioaccumulation in Se-rich regions, are not taken into account by statistically based methods.

With the increasing availability of soil environment-related data, machine learning (ML) technologies are increasingly employed as efficient and high-precision tools to uncover potential patterns among high-dimensional data [12,43]. From a data-driven perspective, ML spatial prediction techniques have shown good performance in handling complex nonlinear problems in the environmental domain. ML spatial prediction techniques are based on the correlation between the environmental covariates and target variable [3]. Using interpretable ML, the inclusion of spatially correlated covariates with location information, such as elevation, distance from the main road, nearest observations, and their distances to the prediction, further enhances the predictive performance and interpretability of the model [30]. This is a promising tool for exploring the complex geochemical relation between karst soil and heavy metals.

Herein, ML is used to study the bioaccumulation ability of Cd in Se-rich karst regions and quantify the contributions of driving factors to the Cd bioaccumulation ability of tobacco. This study aims to (1) establish soil-tobacco systems using XGBoost to predict the Cd bioaccumulation ability of tobacco, (2) quantify the relative contributions of key drivers affecting the Cd bioaccumulation in tobacco, and (3) identify the ecological tobacco-growing zone.

2. Methods and materials

2.1. Study area

Zunyi City (27° 8′-29° 12′ N, 105°36′-108° 13′ E) is located in the central region of Guizhou Province. Its karst landforms are widely distributed, accounting for ~75% of the total land area of the city. It is widely recognized as a high-geochemical background region. Zunyi City is located in the subtropical highland humid monsoon region and characterized by distinct seasons throughout the year, with abundant rainfall and high temperatures occurring in monsoon. The special topographic and climatic conditions and high land utilization have made Zunyi City the primary region for tobacco cultivation in Guizhou Province.

2.2. Sample collection and analysis

A total of 686 samples, including 365 paired rhizospheric soil-grown tobacco samples and 321 topsoil samples, were collected from Zunyi City (Fig. 1). For each sample, the corresponding soil type, lithology, cultivation time, and distance from the main road were recorded. The main physicochemical properties of the soil samples (i.e., pH and organic matter, N, P, and K contents) were determined according to the NF ISO 10694 standard, and the concentrations of Cd and Se in soil and tobacco were measured via flame atomic absorption spectrometry. Furthermore, to ensure the reliability of results, blank measurements were performed after every 20 samples, and each blank sample measurement was performed in triplicate.

2.2.1. Environmental covariate collection

Environmental covariates that considerably affect tobacco growth were measured to better understand the Cd bioaccumulation ability of tobacco and key environmental drivers of the bioaccumulation process. The annual average temperature (**Temp**), annual average precipitation (**Prec**), annual evapotranspiration (**Evp**), annual average relative humidity (**Hum**), annual average ground temperature (**GT**), annual sunshine hour (**Ssh**), and normalized difference vegetation index (**NDVI**) in Zunyi City were obtained from the meteorological monitoring database (https://www.resdc.cn/). The soil cation exchange capacity (**CEC**) was obtained from the Second National Soil Survey in China (http://www.soil.csdb.cn/).

Although elevated soil Cd contents in the karst region are primarily due to geogenic inputs from their high-geochemical background, the anthropogenic input of Cd cannot be overlooked [23]. Specifically, sampling sites closer to the main road are more susceptible to anthropogenic disturbances, such as intensive industrial activities and increased vehicular exhaust emissions, than those farther from the main road. Furthermore, fertilizers are an important source of Cd, which gradually accumulates in the soil over time with continued cultivation. Therefore, the distance of the sampling site from the main road (**Distance**) and cultivation time (**CT**) were used as indirect indicators of traffic emissions and fertilizer application to represent the impact of

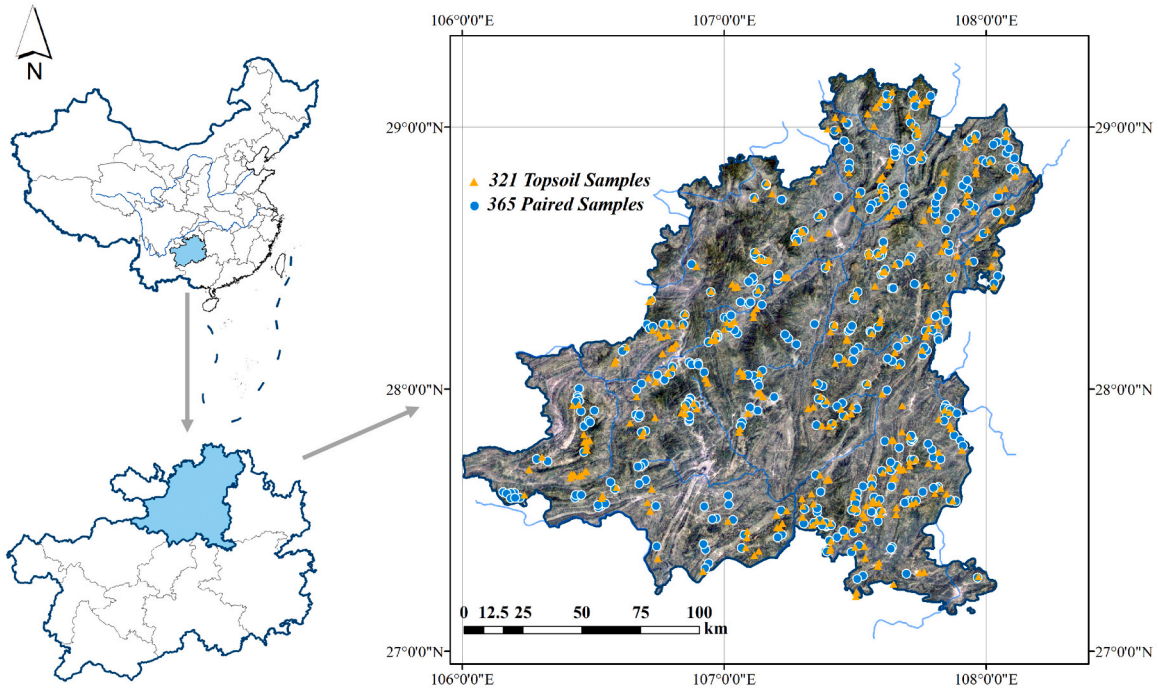


Fig. 1. Geographical location of the study area and layout of sampling points.

anthropogenic activities on the soil Cd content. Finally, 686 data samples with 20 environmental covariates in four categories were collected (Table S1).

2.3. Data preprocessing and feature engineering

Based on the collection of 365 paired samples, the bioaccumulation factor (BAF) was used to quantitatively analyze the Cd bioaccumulation ability of tobacco, which was calculated using Eq. (1).

$$BAF = Cd_{tobacco}/Cd_{soil} \tag{1}$$

where $Cd_{tobacco}$ and Cd_{soil} are the contents of Cd in tobacco and corresponding rhizosphere soil, respectively.

After the logarithmic transformation of the BAF, descriptive statistics was performed on the continuous variables using Python 3.9 software to determine the general status of Cd contamination and other environmental covariates. In general, datasets often include extreme values that lie outside the expected range and exhibit distinct characteristics compared to the majority of the data. These exceptional data points, commonly referred to as outliers, can significantly impact machine learning models, making it crucial to exclude them for improved performance [7]. Data exceeding triple standard deviations away from the mean are considered rare and are identified as outliers. We subsequently exclude these outliers using Eq. (2).

$$P(|x_i - \mu| > 3\sigma) \leq 0.003 \tag{2}$$

where x_i is the bioaccumulation factor of the i-th sample, μ and σ are the mean and standard deviation of bioaccumulation factor.

Target encoding (**TE**) is a highly effective method for encoding categorical columns and occupies only one feature space. It avoids the proliferation of model feature dimensions caused by traditional coding methods(e.g., one-hot encoding and ordinal encoding), while retaining more label-related information. Makes it easier for XGBoost to mine for potential relationships between the influencing factors and heavy metal accumulation in tobacco from the soil. Herein, the categorical variables were encoded using the target encoding technique, wherein each categorical value is replaced by the corresponding average of the dependent variable for that category. Additionally, data smoothing techniques were performed to address potential bias resulting from uneven categories of classification variables. More information about TE is given in **Text S1**.

2.4. Model development

XGBoost is a scalable end-to-end ensemble learning technique for tree boosting that builds on previous ideas in gradient boosting and can implement various types of gradient-boosting trees [2]. XGBoost has attracted considerable attention in the field of spatial prediction because of its remarkable ability to deal with high-dimensional data [28]. Because of this unique feature of XGBoost, variable selection was not performed to eliminate covariates that may exhibit high correlation or collinearity [13,41]. Instead, all available covariates associated with heavy metals were included. Herein, the XGBoost algorithm, which was optimized for hyperparameters using the GridSearch method, was used to predict the Cd bioaccumulation ability of tobacco grown in typical karst tobacco-growing soil. Additionally, to better understand the key environmental drivers controlling the bioaccumulation of Cd in tobacco in the karst region, the relative contributions of predictors were quantitatively evaluated using the feature importance (FI) function in the XGBoost algorithm. Furthermore, a threshold was introduced for screening out important factors with FIs of $> 5\%$. The XGBoost algorithm was implemented based on the Scikit-Learn API (https://scikit-learn.org/) in Python 3.9 software.

2.5. Model performance evaluations

The regression model and loss function were specified as the learning task and objective, respectively, and the coefficient of determination (R^2) and root-mean-square error (RMSE) were used to evaluate the performance of the model. The formulas for calculating R^2 and RMSE are shown in Eqs. (3) and (4), respectively.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_i^{ture} - y_i^{pred}\right)^2}{\sum_{i=1}^{N}\left(y_i^{ture} - \bar{y}\right)^2} \tag{3}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i^{ture} - y_i^{pred})^2} \quad (4)$$

where y_i^{ture} and y_i^{pred} are the true value of the adsorption ability and predicted value of the i-th sample, respectively, $\bar{y}$ is the mean value of the experimentally measured adsorption ability, and N is the sample size.

To ensure the model's robustness and generalization capabilities, the dataset, comprising 356 paired samples, was split into training and testing sets using the CVT (training with cross-validation (CV), and testing) method as proposed by Zhu et al. [44] with an 8:2 ratio. Subsequently, fivefold cross-validation was performed on the training set, and the hyperparameters were tuned based on the results of fivefold cross-validation. The generalization performance of the model was evaluated using the validation set. Furthermore, 321 topsoil samples, which served as a field validation combined with 356 paired samples were used for digital mapping of status of Cd contamination.

3. Results and discussion

3.1. Descriptive statistics of elemental contents and physicochemical properties in soil and tobacco samples

The descriptive statistics of the soil samples in the study area is shown in Table 1. Overall, the soil pH ranges from 3.75 to 8.69, with a mean value of 6.46. The SOM content ranges from 1.27% to 17.34%, with an average of 3.47%. The highest difference in the SOM content between soil samples is approximately 17-fold. The N, P, and K contents in the soil are 0.2%– 3.7%, 0.2%– 2.1%, and 5.2%– 55.2%, respectively. The average N, P, and K contents in the soil are 1.8%, 0.8%, and 1.96%, respectively. The coefficient of variance (CV) values of soil pH and SOM contents are relatively low, that is, 0.16 and 0.41, respectively, indicating the absence of notable spatial heterogeneity. Similar patterns are observed for the N, P, and K contents in soil samples, with CV values of 0.28, 0.38, and 0.40, respectively.

In terms of heavy metal contents, the Cd concentration in tobacco-growing soil (Cd_{soil}) ranges from 0.18 to 3.81 mg/kg, with an average value of 0.65 mg/kg. According to the GB15618–2018 standard on the screening levels for soil contamination of agricultural land [26], the point exceedance rate for Cd is 90.17%. In comparison, the Cd content in tobacco ($Cd_{tobacco}$) ranges from 0.43 to 17.77 mg/kg, with an average of 4.21 mg/kg. The BAFs of Cd (BAF_{Cd}) between soil and tobacco samples range from 0.71 to 38.56, with an average value of 8.02. Similar to the distribution pattern of the potentially toxic element Cd, Se also exhibits a notable enrichment pattern (Fig. S1). The soil Se content (Se_{soil}) ranges from 0.22 to 2.90 mg/kg, with an average value of 0.58 mg/kg. The Se-enrichment rate in the soil samples from the study area is 75.84%. In contrast to soil physicochemical properties, soil and tobacco exhibit a notable heterogeneity in the Cd content, with CV values of 0.71 and 0.74, respectively. Additionally, the calculated CV value of the BAF_{Cd} between soil and tobacco is 0.81, which is the highest among all CV values for the samples. This indicates that the Cd bioaccumulation process in tobacco may be influenced by multiple factors, such as point source contamination and regional disparities. Furthermore, the CV value of Se_{soil} is high, exhibiting a certain degree of coupling with Cd_{soil} and $Cd_{tobacco}$. This indicates that a change in Se_{soil} may partially explain the geochemical processes of Cd enrichment in tobacco with a high-geochemical background, which will be discussed in Section 3.4. Descriptive statistics for other environmental covariates is provided in Table S2.

3.2. Evaluation and application of the XGBoost algorithm

After optimizing the hyperparameters via fivefold cross-validation, XGBoost was used to establish the relationship between BAF_{Cd} and 20 environmental covariates to predict BAF_{Cd} and analyze the contribution of key driving factors.

The predictive performance of XGBoost is shown in Fig. S2. XGBoost demonstrated strong learning ability, achieving an R^2 of 0.86 and an RMSE of 0.012 on the training set. These results indicate that the model effectively captures the potential relationships between 20 environmental covariates and BAF_{Cd}. However, a slightly lower R^2 of 0.44 is achieved when using the validation set, indicating a potential risk of overfitting in the model. This trend may be related to the conflict between the high data dimensionality and low data volume, resulting in the model converging to a local optimal solution instead of a global one [6]. Furthermore, the intricate coupling mechanisms between soil heavy metals and tobacco in the high-geochemical background region reduced the explanatory power of the model for the target variable. Nonetheless, the RMSE of XGBoost is lower than the standard deviation (std) of the measured values (RMSE and std of 0.06 and 3.15, respectively). This indicates that the generalization ability of the model is satisfactory and XGBoost produces better predictions than the models built directly using the measured values alone [35]. Additionally, the slight difference between the predicted (8.1 mg/kg) and observed (8.2 mg/kg) mean values further confirms the effectiveness of the model.

3.3. Influence of key factors on the bioaccumulation of Cd in tobacco in Se-rich area

The bioaccumulation of Cd in tobacco is affected by various factors, including natural and anthropogenic sources of Cd. Previous research has predominantly investigated the effects of different factors on the bioaccumulation of Cd in tobacco in non-Se-rich regions. Given the special occurrence of Se anomalies in the study area, this study specifically addresses the factors that affect Cd bioaccumulation in tobacco in high-Se geochemical background areas.

As shown in Fig. 2, the overall contributions of the four categories to Cd bioaccumulation in tobacco follow the order: regional geochemical parameters > soil physicochemical properties > tobacco cultivation conditions > anthropogenic activities. In terms of regional geochemical characteristics with FI > 5%, Cd_{soil} plays a critical role in the bioaccumulation of Cd in tobacco, which explain 35.4% of the variation in

Table 1
Descriptive statistics of 356 paired data.

	Cd_{soil}	$Cd_{tobacco}$	Se_{soil}	BAF_{Cd}	pH	SOM	N	P	K
	mg/kg				-	%			
Max	3.81	17.77	2.90	38.56	8.69	17.34	0.37	0.21	5.52
Min	0.18	0.43	0.22	0.71	3.75	1.27	0.02	0.02	0.52
Mean	0.65	4.21	0.58	8.02	6.46	3.47	0.18	0.08	1.96
Medium	0.53	3.37	0.50	6.46	6.5	3.22	0.18	0.08	1.87
Std	0.46	3.15	0.33	6.46	1.04	1.43	0.05	0.03	0.79
CV(-)	0.71	0.74	0.56	0.81	0.16	0.41	0.28	0.38	0.40
Point over standard rate of Cd				90.17%		Se-rich soil rate			75.84%

The point over standard rate of Cd and Se-rich soil rate are calculated based on Soil environmental quality risk control standard for soil contamination of agricultural land(GB 15618–2018) and Delineation and identification of naturally selenium-rich land(DD 2019–10).

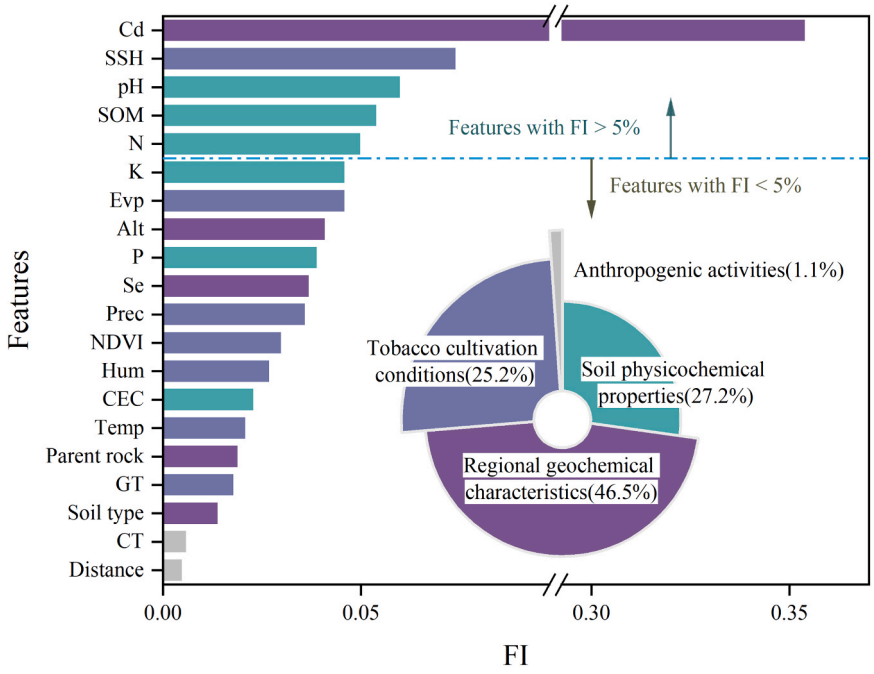


Fig. 2. Relative contributions of 20 key environmental covariates to Cd bioaccumulation in tobacco.

the BAF_{Cd}. This is primarily controlled by the biogeochemical behavior of Cd in crops. Specifically, the presence of heavy metals in tobacco is predominantly attributed to its natural ability to accumulate these contaminants from soils [18]. It has been observed that tobacco grown in soil with high concentrations of heavy metals tends to accumulate elevated levels of these elements [36]. Therefore, Cd_{soil} commonly explains 94%– 97% of the overall variation in non-Se-rich areas [24].

However, herein, the explanatory power of Cd_{soil} is decreased. Furthermore, the influence of soil type and lithology on the variation in BAF_{Cd} is limited, explaining only 2.32% and 2.11% of the total variation, respectively. This differs from the findings of Hu and Cheng, [9], who reported that the soil type is one of the main factors controlling the Cd content in the Pearl River Delta (PRD) in China. This discrepancy may be because of the constraint of Se on Cd in high-geochemical background areas. The results indicate that Se_{soil} has an undeniable effect on the bioaccumulation of Cd in tobacco, accounting for 3.70% of the overall variation in the BAF_{Cd}. Se_{soil} is largely determined by the Se content in the parent material and parent rock. Compared to the yellow-brown earth in the PRD, the red-yellow earth in southwestern China is enriched with higher levels of Se (0.04–2.55 mg/kg) [21,29].

The relation between Se_{soil} and BAF_{Cd} is shown in Fig. S3. The Spearman correlation analysis results show a negative correlation with low strength ($R^2 = -0.16$; $P < 0.01$) between Se_{soil} and BAF_{Cd}. Additionally, the scatter diagram (Fig. 3**a**) shows that the range of BAF_{Cd} greatly varies between 0.71 and 38.56 when Se_{soil} is relatively low (Se_{soil} < 0.85 mg/kg), indicating that Se_{soil} has a limited effect on the bioaccumulation of Cd in tobacco. By contrast, as Se_{soil} increases, the BAF_{Cd} values do not considerably vary (all <20), indicating that a higher Se_{soil} may inhibit the bioaccumulation of Cd in tobacco. This finding is consistent with the literature, in which it was proposed that the formation of the Cd-Se complex in the roots of plants hinders further metabolism and transformation of Cd or Se to aerial parts of plants [22, 40]. Furthermore, the results of the partial dependence analysis are shown in Fig. 3**b**. BAF_{Cd} first increases and then decreases with an increase in Se_{soil}, consistent with the findings of Guo et al. [5], who reported that low Se_{soil} increases the bioaccumulation of Cd in crops, but BAF_{Cd} in all parts of the crop is at the lowest level when the molar ratio of Se/Cd is > 1. Using ML, the threshold for the mutual suppression of Se and Cd at a regional scale was determined. The findings indicate that within a Cdsoil range of 0.18–3.81 mg/kg, elevated Se_{soil} plays an undeniable role in the bioaccumulation of Cd in tobacco, and the highest Cd bioaccumulation ability is observed when Se_{soil} is 0.8 mg/kg.

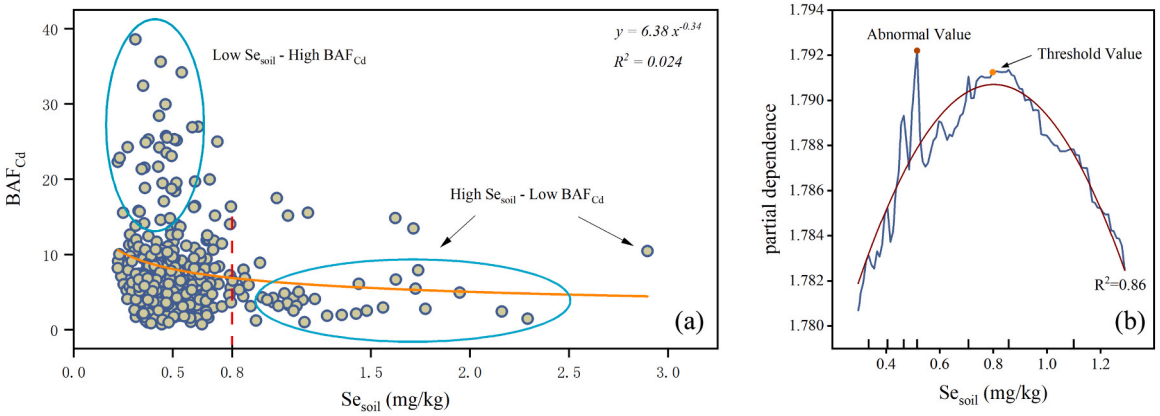


Fig. 3. Correlation between BAF_{Cd} and Se_{soil} (a); Partial dependence plot of Se_{soil} (b).

Among the soil physicochemical properties with FI > 5%, the soil pH, SOM content, and N content are the key determinants controlling BAF_{Cd}, accounting for 6%, 5.4%, and 5% of the total variation in BAF_{Cd}, respectively. Compared to Cd_{soil}, the influence of soil physicochemical properties on the change in BAF_{Cd} is only half as significant. This may be attributed to the negative feedback between these factors and BAF_{Cd}. For example, the extensive weathering and dissolution of carbonate rocks in the study area increase soil pH, whereas heavy precipitation and intense leaching decrease the soil nutrient content. This results in soil impoverishment, which profoundly affects the bioaccumulation of Cd in tobacco. As the soil pH increases, the adsorption of Cd^{2+} by soil colloids also increases, reducing the Cd^{2+} availability in soil and decreasing the accumulation of Cd in tobacco [16].

Moreover, the various functional groups in SOM (e.g., hydroxyl, carboxyl, and phenolic groups) solubilize Cd in soils via complexation, forming Cd-dissolved organic matter complexes, and ligand-assisted dissolution of solid binding phases (i.e., metal oxyhydroxides) and consequently reduce the bioavailability of Cd in soil solution [4,33]. Additionally, the content of essential nutrients needed for tobacco growth considerably affects Cd accumulation in crops by affecting gene expression, enzyme activity and subcellular distribution in plant cells [16]. For example, the overexpression of genes encoding pectin methylesterase in the cell wall promotes pectin demethylation, which conversely promotes Cd adsorption by cell walls [27].

Although the covariates representing tobacco cultivation conditions, such as Evp, Alt, and Prec, have FI values of < 5%, these factors may still have some influence on the bioaccumulation of Cd in tobacco. The Evp, Alt, and Prec explain 12.3% and Ssh explains 7.4% of the overall variation in BAF_{Cd}, indicating that the growth conditions of tobacco also affect its ability to bioaccumulate heavy metals to some extent. This viewpoint is supported by the study of Li et al. [17], who demonstrated that the enrichment of heavy metals is considerably correlated with geoclimatic conditions such as altitude, precipitation, and temperature. Furthermore, the influence of these factors may be twofold. On the one hand, the temperature decreases with an increase in elevation, decreasing the decomposition rate of SOM, which further decreases the Cd leaching content in the soil [31]. On the other hand, the complex topography in karst areas leads to local microclimate anomalies, increasing the leaching and migration frequency of Cd in the soil because of the high amount of rainfall [17]. Additionally, these cultivation conditions can affect the distribution of heavy metals in the topsoil by influencing soil erosion and redistribution, thus affecting the bioaccumulation of Cd in tobacco [42].

In the results of this study, the FIs of CT and Distance, which represent the effect of anthropogenic activity on soil Cd contents, are not significant and explain only 0.6% and 0.5% of the total variation in BAF_{Cd}, respectively. This may be because of the strong influence of regional geochemical characteristics and soil physicochemical properties on BAF_{Cd}, which overshadow the influence of human activities. Moreover, the special karst landforms within the region, such as peaks, gorges, and underground water systems, have constrained the development of transportation and industrial activities in the area, resulting in relatively weak anthropogenic impacts [10].

3.4. Implications for the identification of tobacco cultivation areas

Using 365 and 321 data points, the spatial distribution of BAF_{Cd} in Zunyi City was predicted via XGBoost. As shown in Fig. 4, the distribution of the Cd bioaccumulation ability of tobacco is clearly uneven among the study area. High-BAF_{Cd} areas are mainly distributed in a few villages in the south, central, and northeast of Zunyi City. Herein, the effective regulation of Se_{soil} considerably reduced the bioaccumulation of Cd in tobacco. This provides favorable cultivation conditions for the development of the tobacco industry in Zunyi City. Consequently, a mirror symmetric relationship is observed between the spatial distributions of Se_{soil} and BAF_{Cd}. That is, regions with high Se_{soil} typically exhibit lower BAF_{Cd} values (BAF_{Cd} < 20), particularly in the northern and southeastern areas of Zunyi City (Fig. 4 and **S2**).

The spatial correlation between Se_{soil} and BAF_{Cd} was determined using the bivariate local indicators of spatial association analysis (LISA) to assess the potential interdependence and heterogeneity between the two variables (More information on LISA is given in **Text S2**). The spatial correlation patterns of these two distinct variables guided the identification of soil control areas for tobacco cultivation in Zunyi City. Different from the traditional LISA analysis (Fig. S4), significant regions with high BAF_{Cd}-high Se_{soil} and high BAF_{Cd}-low Se_{soil} were delineated as

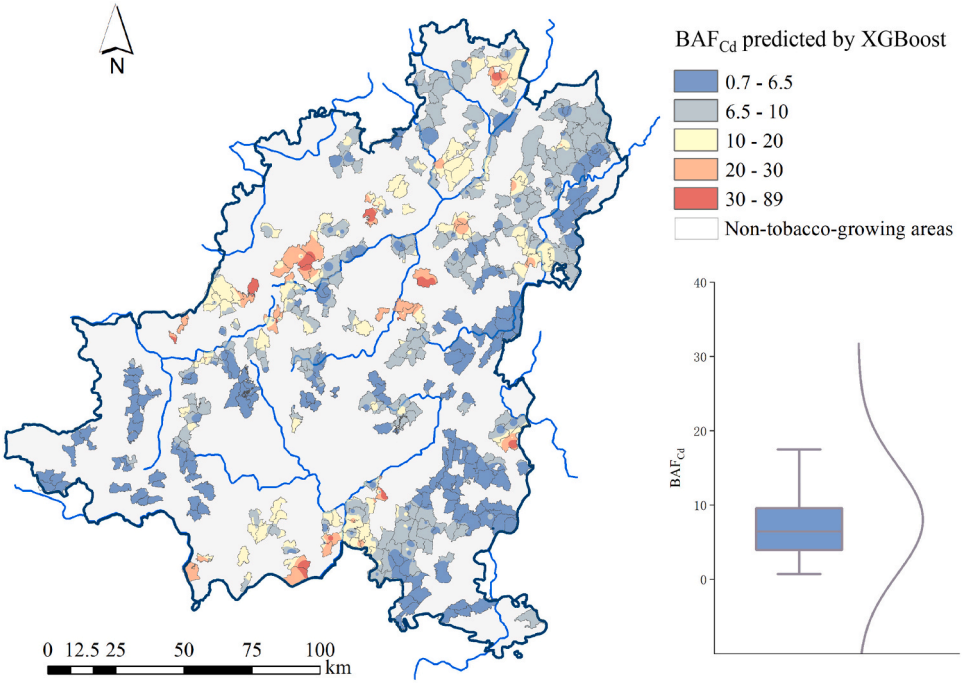


Fig. 4. Spatial distribution of BAF_{Cd} predicted by XGBoost.

priority areas for heavy metal remediation and low BAF_{Cd}-high Se_{soil} and low BAF_{Cd}-low Se_{soil} were integrated as areas for ecological tobacco cultivation. Tobacco grown in regions with high BAF_{Cd}-low Se_{soil} carries a risk of bioaccumulating Cd, and the suppressive effect of Se_{soil} is limited. Consequently, tobacco grown in these regions is prone to bioaccumulating excessive levels of Cd, posing a threat to public health. Furthermore, in regions with high BAF_{Cd}-high Se_{soil}, excessive Se_{soil} may not effectively inhibit the bioaccumulation of Cd in tobacco and could potentially result in Se contamination [21]. Therefore, implementation of remediation measures for heavy metals should be a priority. Conversely, regions with low BAF_{Cd}-high Se_{soil}, and those with low BAF_{Cd}-low Se_{soil}, are have a lower risk to public health if appropriate management practices are implemented. Therefore, these regions are more suitable for ecological tobacco cultivation.

The regionalization results are shown in Fig. 5. The priority heavy metal remediation zones are mainly concentrated in the eastern parts of Fenggang County, the central part of Meitan County, the southeast part of Yuqing County, the junction of Tongzi County and Zheng'an County, and the junction of Suiyang County and Fenggang County. In addition, there are several spatial clusters distributed in the central part of Daozhen County, the central and northern parts of Wuchuan County, and the junction of Huichuan District and Tongzi County. Remarkably, Bozhou District, Suiyang County, and Huairen County exhibit localized spatial aggregations characterized by a few points, possibly attributable to point source pollution. It is noteworthy that Honghuagang District, Bozhou District, and Huairen County serve as primary industrial zones within Zunyi City. The emissions released from these concentrated industrial establishments plausibly account for the limited occurrence of point spatial aggregations in those specific regions [36].

The regionalization results are shown in Fig. 5. The priority heavy metal remediation zones are mainly concentrated in the eastern regions of Fenggang County, central region of Meitan County, southeast region of Yuqing County, junction of Tongzi County and Zheng'an County, and junction of Suiyang County and Fenggang County. Additionally, there are several spatial clusters distributed in the central region of Daozhen County, central and northern regions of Wuchuan County, and junction of Huichuan District and Tongzi County. Remarkably, Bozhou District, Suiyang County, and Huairen County exhibit localized spatial aggregations characterized by a few points, possibly due to point source pollution. Notably, Honghuang District, Bozhou District, and Huairen County are the primary industrial zones within Zunyi City. The emissions released from industrial establishments concentrated in these regions plausibly account for the limited occurrence of point spatial aggregations in these regions [36].

In contrast to the distribution of priority governance zones, the ecological tobacco-growing zone avoid the industry-concentrated zones in Zunyi City and are mainly concentrated in the central areas of Wuchuan County, Daozhen County, Zheng'an County and borders of Suiyang County, Huichuan District, and Tongzi County. A few spatial clusters are distributed in the western and southern regions of Fenggang County, central-southern region of Meitan County, eastern region of Bozhou District, and eastern region of Yuqing County.

4. Conclusions

Regional disparities and changes in environmental conditions considerably change the Cd bioaccumulation ability of tobacco in areas with high-Se geochemical backgrounds. Feature importance-analysis results indicate that regional geochemical characteristics (e.g., elevation and soil Cd and Se contents) have the highest influence on BAF_{Cd}, followed by physicochemical properties of soil (e.g., soil pH and SOM content) and tobacco cultivation conditions (e.g., Ssh and Perc). The effect of anthropogenic activities is relatively insignificant. From a data-driven perspective, elevated Se_{soil} undeniably affects the bioaccumulation of Cd in tobacco and the mutual suppression effects of Se and Cd become more progressively pronounced when Se_{soil} is 0.8 mg/kg. Compared to non-Se-rich regions, this mutual suppression effect is responsible for the relatively low explanation rate of Cd_{soil}, soil type, and lithology. Furthermore, the spatial distribution of BAF_{Cd} shows considerable heterogeneity under the intricate biogeochemical behavior between soil Cd and tobacco in the karst region. ML and LISA analysis results show that tobacco cultivated in the central, northeast, and southeast Zunyi City carries a low risk of soil Cd contamination.

The proposed method, which integrates an ML model and geostatistical analysis, is an effective tool to quantify the contributions of key factors in the Cd bioaccumulation process of tobacco and identify the ecological tobacco-growing zones in high-geochemical background areas. This methodologically positive attempt is intended to provide researchers and decision-makers with a basis for the scientific management of tobacco cultivation lands in the karst regions.

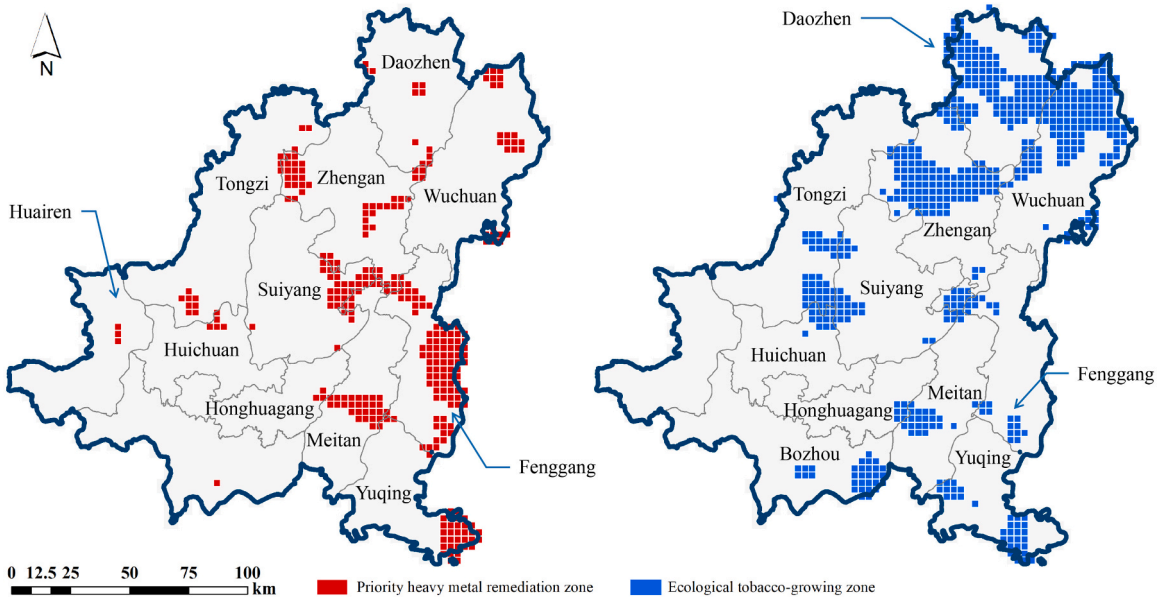


Fig. 5. Spatial distributions of ecological tobacco-growing zone and priority heavy metal remediation zones.

Environmental implication

Cadmium (Cd) exhibits higher mobility and toxicity to living organisms than other hazardous heavy metals. Tobacco is one of the main economic crops grown in high-geochemical background areas of southwest China. Tobacco grown in high-geochemical background areas tends to bioaccumulate more Cd in its leaves, posing a threat to human health. This study quantifies the key driving factors affecting Cd bioaccumulation in tobacco from a data-driven perspective. XGBoost and bivariate local indicators of spatial association analysis are simultaneously used to provide comprehensive insights to identify optimal regions for tobacco cultivation.

CRediT authorship contribution statement

Zilun Gou: Writing – original draft, Visualization, Data curation. **Chengshuai Liu:** Conceptualization, Writing - review & editing, Funding acquisition. **Meng Qi:** Writing – review & editing, Methodology. **Yajing Qu:** Data curation, Visualization. **Wenhao Zhao:** Methodology, Data curation. **Yi Sun:** Data curation. **Yajing Qu:** Data curation. **Jin Ma:** Conceptualization, Writing - review & editing, Funding acquisition.

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.

Data Availability

The authors do not have permission to share data.

Acknowledgments

This work was supported by the National Natural Science Foundations of China (42025705, 42177221) and the Guizhou Province High-level Talent Project (GCC[2022]002-1).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jhazmat.2023.132910.

References

[1] Cao, J.F., Li, C.F., Wu, Q.Y., Qiao, J.M., 2020. Improved mapping of soil heavy metals using a vis-NIR spectroscopy index in an agricultural area of Eastern China. IEEE Access 8, 42584–42594. https://doi.org/10.1109/ACCESS.2020.2976902.

[2] Chen, T.Q., Guestrin, C., 2016. Xgboost: a scalable tree boosting system. Proc 22nd ACM Sigkdd Int Conf Knowl Discov Data Min ACM 785–794. https://doi.org/10.1145/2939672.2939785.

[3] Gao, B.B., Stein, A., Wang, J.F., 2022. A two-point machine learning method for the spatial prediction of soil pollution. Int J Appl Earth Obs Geoinf 108, 102742. https://doi.org/10.1016/j.jag.2022.102742.

[4] Gao, X.P., Brown, K.R., Racz, G.J., Grant, C.A., 2010. Concentration of cadmium in durum wheat as affected by time, source and placement of nitrogen fertilization under reduced and conventional-tillage management. Plant Soil 337, 341–354. https://doi.org/10.1007/s11104-010-0531-y.

[5] Guo, Y.K., Mao, K., Cao, H.R., Ali, W., Lei, D., Teng, D.Y., et al., 2021. Exogenous selenium (cadmium) inhibits the absorption and transportation of cadmium (selenium) in rice. Environ Pollut 268, 115829. https://doi.org/10.1016/j.envpol.2020.115829.

[6] Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J Mach Learn Res 3, 1157–1182 https://dl.acm.org/doi/10.5555/944919.944968.

[7] Harris, G.L., Torres, J.A., 2003. Selected laboratory and measurement practices and procedures to support basic mass calibrations. NIST Interag/Intern Rep 6969. https://doi.org/10.6028/NIST.IR.6969-2019.

[8] He, J.Y., Yang, Y., Christakos, G., Liu, Y.J., Yang, X., 2019. Assessment of soil heavy metal pollution using stochastic site indicators. Geoderma 337, 359–367. https://doi.org/10.1016/j.geoderma.2018.09.038.

[9] Hu, Y.N., Cheng, H.F., 2013. Application of stochastic models in identification and apportionment of heavy metal pollution sources in the surface soils of a large-scale region. Environ Sci Technol 47, 3752–3760 https://pubs.acs.org/doi/10.1021/es304310k.

[10] Hu, Y.N., Cheng, H.F., Tao, S., 2016. The challenges and solutions for cadmium-contaminated rice in China: a critical review. Environ Int 92–93, 515–532. https://doi.org/10.1016/j.envint.2021.106749.

[11] Huang, B.Y., Lü, Q.X., Tang, Z.X., Tang, Z., Chen, H.P., Yang, X.P., et al., 2023. Machine learning methods to predict cadmium (Cd) concentration in rice grain and support soil management at a regional scale. Fundam Res. https://doi.org/10.1016/j.fmre.2023.02.016.

[12] Janet, J.P., Kulik, H.J., Morency, Y., Caucci, M.K., 2020. Machine Learning in Chemistry. American Chemical Society. https://doi.org/10.1021/acs.infocus.7e4001.

[13] Jia, X.L., Hu, B.F., Marchant, B.P., Zhou, L.Q., Shi, Z., Zhu, Y.W., 2019. A methodological framework for identifying potential sources of soil heavy metal pollution based on machine learning: a case study in the Yangtze Delta, China. Environ Pollut 250, 601–609.

[14] Jiang, Z.H., Liu, H.Y., Wang, H.Y., Peng, J., Meersmans, J., Green, S.M., et al., 2020. Bedrock geochemistry influences vegetation growth by regulating the regolith water holding capacity. Nat Commun 11, 2392. https://doi.org/10.1038/s41467-020-16156-1.

[15] Li, C., Zhang, C.S., Yu, T., Liu, X., Yang, Y.Y., Hou, Q.Y., et al., 2022. Use of artificial neural network to evaluate cadmium contamination in farmland soils in a karst area with naturally high background values. Environ Pollut 304, 119234. https://doi.org/10.1016/j.envpol.2022.119234.

[16] Li, H., Luo, N., Li, Y.W., Cai, Q.Y., Li, H.Y., Mo, C.H., et al., 2017. Cadmium in rice: transport mechanisms, influencing factors, and minimizing measures. Environ Pollut 224, 622–630. https://doi.org/10.1016/j.envpol.2017.01.087.

[17] Li, X.Y., Geng, T., Shen, W.J., Zhang, J.R., Zhou, Y.Z., 2021. Quantifying the influencing factors and multi-factor interactions affecting cadmium accumulation in limestone-derived agricultural soil using random forest (RF) approach. Ecotoxicol Environ Saf 209, 111773. https://doi.org/10.1016/j.ecoenv.2020.111773.

[18] Lin, B.S., Gao, H.J., Lai, H.M., Li, C.H., Wang, Q., 2016. Characterization of heavy metals in soils from typical tobacco cultivated areas, China. Environ Prog Sustain Energy 36, 483–488. https://doi.org/10.1002/ep.12505.

[19] Liu, H.L., Wang, X.Q., Zhang, B.M., Han, Z.X., Wang, W., Chi, Q.H., et al., 2021. Concentration and distribution of selenium in soils of mainland China, and implications for human health. J Geochem Explor 220, 106654. https://doi.org/10.1016/j.gexplo.2020.106654.

[20] Liu, H.W., Wang, H.Y., Zhang, Y., Yuan, J.M., Peng, Y.D., Li, X.C., et al., 2018. Risk assessment, spatial distribution, and source apportionment of heavy metals in Chinese surface soils from a typically tobacco cultivated area. Environ Sci Pollut Res 25, 16852–16863. https://doi.org/10.1007/s11356-018-1866-9.

[21] Liu, Q.Y., Wu, Y.H., Zhou, Y.Z., Li, X.Y., Yang, S.H., Chen, Y.X., et al., 2021. A novel method to analyze the spatial distribution and potential sources of pollutant combinations in the soil of Beijing urban parks. Environ Pollut 284, 117191. https://doi.org/10.1016/j.envpol.2021.117191.

[22] Liu, W.X., Shang, S.H., Feng, X., Zhang, G.P., Wu, F.B., 2015. Modulation of exogenous selenium in cadmium-induced changes in antioxidative metabolism, cadmium uptake, and photosynthetic performance in the 2 tobacco genotypes differing in cadmium tolerance. Environ Toxicol Chem 34, 92–99. https://doi.org/10.1002/etc.2760.

[23] Liu, Y.Z., Xiao, T.F., Perkins, R.B., Zhu, J.M., Zhu, Z.J., Xiong, Y., et al., 2017. Geogenic cadmium pollution and potential health risks, with emphasis on black shale. J Geochem Explor 176, 42–49. https://doi.org/10.1016/j.gexplo.2016.04.004.

[24] Lu, X., Zhang, D., Ugurlu, A., Chen, Y.L., Proshad, R., 2021. Bioaccumulation of Cadmium in Nicotiana tabacum L. (tobacco) characterized by soil properties: a case study in the Sichuan basin, China. Anal Lett 54, 2883–2894. https://doi.org/10.1080/00032719.2021.1900215.

[25] Marano, K.M., Naufal, Z.S., Kathman, S.J., Bodnar, J.A., Borgerding, M.F., Garner, C.D., et al., 2012. Cadmium exposure and tobacco consumption: biomarkers and risk assessment. Regul Toxicol Pharmacol 64, 243–252. https://doi.org/10.1016/j.yrtph.2012.07.008.

[26] MEE, 2018. Ministry of Ecology and Environment of the People's Republic of China. Soil environmental quality Risk control standard for soil contamination of agricultural land. GB 15618–2018.

[27] Mei, S.N., Lin, K.N., Williams, D.V., Liu, Y., Dai, H.X., Cao, F.B., 2022. Cadmium accumulation in cereal crops and tobacco: a review. Agronomy 12, 1952. https://doi.org/10.3390/agronomy12081952.

[28] Pan, B.Y., 2018. Application of XGBoost algorithm in hourly PM2.5 concentration prediction. IOP Conf Ser: Earth Environ Sci 113, 012127. https://doi.org/10.1088/1755-1315/113/1/012127.

[29] Pan, Z.P., He, S.L., Li, C.J., Men, W., Yan, C.Z., Wang, F., 2017. Geochemical characteristics of soil selenium and evaluation of Se-rich land resources in the central area of Guiyang City, China. Acta Geochim 36, 240–249. https://doi.org/10.1007/s11631-016-0136-0.

[30] Sekulić, A., Kilibarda, M., Heuvelink, G.B.M., Nikolić, M., Bajat, B., 2020. Random forest spatial interpolation. Remote Sens 12, 1687. https://doi.org/10.3390/rs12101687.

[31] Shi, T.R., Zhang, Y.Y., Gong, Y.W., Ma, J., Wei, H.Y., Wu, X., et al., 2019. Status of cadmium accumulation in agricultural soils across China (1975–2016): from temporal and spatial variations to risk assessment. Chemosphere 230, 136–143. https://doi.org/10.1016/j.chemosphere.2019.04.208.

[32] Wang, Y.Z., Yu, T., Yang, Z.F., Bo, H.Z., Lin, Y., Yang, Q., et al., 2021. Zinc concentration prediction in rice grain using back-propagation neural network

based on soil properties and safe utilization of paddy soil: a large-scale field study in Guangxi, China. Sci Total Environ 798, 149270. https://doi.org/10.1016/j.scitotenv.2021.149270.

[33] Welikala, D., Robinson, B.H., Moltchanova, E., Hartland, A., Lehto, N.J., 2021. Soil cadmium mobilisation by dissolved organic matter from soil amendments. Chemosphere 271, 129536. https://doi.org/10.1016/j.chemosphere.2021.129536.

[34] Wen, Y.B., Li, W., Yang, Z.F., Zhang, Q.Z., Ji, J.F., 2020. Enrichment and source identification of Cd and other heavy metals in soils with high geochemical background in the karst region, Southwestern China. Chemosphere 245, 125620. https://doi.org/10.1016/j.chemosphere.2019.125620.

[35] Were, K., Bui, D.T., Dick, Ø.B., Singh, B.R., 2015. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecol Indic 52, 394–403. https://doi.org/10.1016/j.ecolind.2014.12.028.

[36] Wu, H.W., Liu, Q.Y., Ma, J., Liu, L., Qu, Y.J., Gong, Y.W., et al., 2020. Heavy Metal (loids) in typical Chinese tobacco-growing soils: Concentrations, influence factors and potential health risks. Chemosphere 245, 125591. https://doi.org/10.1016/j.chemosphere.2019.125591.

[37] Xiao, N.C., Wang, F.P., Tang, L.B., Zhu, L.L., Song, B., Chen, T.B., 2022. Recommended risk screening values for Cd in high geological background area of Guangxi, China. Environ Monit Assess 194, 202. https://doi.org/10.1007/s10661-022-09802-2.

[38] Yang, Y.J., Xiong, J., Chen, R.J., Fu, G.F., Chen, T.T., Tao, L.X., 2016. Excessive nitrate enhances cadmium (Cd) uptake by up-regulating the expression of OsIRT1 in rice (Oryza sativa). Environ Exp Bot 122, 141–149. https://doi.org/10.1016/j.envexpbot.2015.10.001.

[39] Zeraatpisheh, M., Garosi, Y., Reza Owliaie, H., Ayoubi, S., Taghizadeh-Mehrjardi, R., Scholten, T., et al., 2022. Improving the spatial prediction of soil organic carbon using environmental covariates selection: a comparison of a group of environmental covariates. CATENA 208, 105723. https://doi.org/10.1016/j.catena.2021.105723.

[40] Zhang, C., Ge, J., Lv, M.W., Zhang, Q., Talukder, M., Li, J.L., 2020. Selenium prevent cadmium-induced hepatotoxicity through modulation of endoplasmic reticulum-resident selenoproteins and attenuation of endoplasmic reticulum stress. Environ Pollut 260, 113873. https://doi.org/10.1016/j.envpol.2019.113873.

[41] Zhang, H., Yin, A.J., Yang, X.H., Fan, M.M., Shao, S.S., Wu, J.T., et al., 2021. Use of machine-learning and receptor models for prediction and source apportionment of heavy metals in coastal reclaimed soils. Ecol Indic 122, 107233. https://doi.org/10.1016/j.ecolind.2020.107233.

[42] Zhao, W.H., Ma, J., Liu, Q.Y., Dou, L., Qu, Y.J., Shi, H.D., et al., 2023. Accurate prediction of soil heavy metal pollution using an improved machine learning method: a case study in the Pearl River Delta, China. Environ Sci Technol. https://doi.org/10.1021/acs.est.2c07561.

[43] Zhong, S.F., Zhang, K., Bagheri, M., Burken, J.G., Gu, A., Li, B.K., et al., 2021. Machine learning: new ideas and tools in environmental science and engineering. Environ Sci Technol 55, 12741–12754. https://doi.org/10.1021/acs.est.1c01339.

[44] Zhu, J.J., Yang, M.Q., R, Z.Y., 2023. Machine learning in environmental research: common pitfalls and best practices. Environ Sci Technol. https://doi.org/10.1021/acs.est.3c00026.