PLOS | ONE

RESEARCH ARTICLE

A Comparative Assessment of the Influences of Human Impacts on Soil Cd Concentrations Based on Stepwise Linear Regression, Classification and Regression Tree, and Random Forest Models

Lefeng Qiu[1], Kai Wang[2], Wenli Long[3], Ke Wang[4], Wei Hu[1]*, Gabriel S. Amable[5]*

1 Institute of Rural Development, Zhejiang Academy of Agricultural Sciences, Hangzhou, China, **2** School of Marine Sciences, Ningbo University, Ningbo, China, **3** Institute of Digital Agriculture, Zhejiang Academy of Agricultural Sciences, Hangzhou, China, **4** Institute of Remote Sensing and Information System Application, Zhejiang University, Hangzhou, China, **5** Department of Geography, University of Cambridge, Cambridge, United Kingdom

* huwei@mail.zaas.ac.cn (WH); gabriel.amable@geog.cam.ac.uk (GSA)

OPEN ACCESS

Citation: Qiu L, Wang K, Long W, Wang K, Hu W, Amable GS (2016) A Comparative Assessment of the Influences of Human Impacts on Soil Cd Concentrations Based on Stepwise Linear Regression, Classification and Regression Tree, and Random Forest Models. PLoS ONE 11(3): e0151131. doi:10.1371/journal.pone.0151131

Editor: Manuel Reigosa, University of Vigo, SPAIN

Received: October 29, 2015

Accepted: February 24, 2016

Published: March 10, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This study was financially supported by National Natural Science Foundation of China (no. 41401595) and National Science and Technology Support Program (no. 2015BAL02B03).

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Soil cadmium (Cd) contamination has attracted a great deal of attention because of its detrimental effects on animals and humans. This study aimed to develop and compare the performances of stepwise linear regression (SLR), classification and regression tree (CART) and random forest (RF) models in the prediction and mapping of the spatial distribution of soil Cd and to identify likely sources of Cd accumulation in Fuyang County, eastern China. Soil Cd data from 276 topsoil (0–20 cm) samples were collected and randomly divided into calibration (222 samples) and validation datasets (54 samples). Auxiliary data, including detailed land use information, soil organic matter, soil pH, and topographic data, were incorporated into the models to simulate the soil Cd concentrations and further identify the main factors influencing soil Cd variation. The predictive models for soil Cd concentration exhibited acceptable overall accuracies (72.22% for SLR, 70.37% for CART, and 75.93% for RF). The SLR model exhibited the largest predicted deviation, with a mean error (ME) of 0.074 mg/kg, a mean absolute error (MAE) of 0.160 mg/kg, and a root mean squared error (RMSE) of 0.274 mg/kg, and the RF model produced the results closest to the observed values, with an ME of 0.002 mg/kg, an MAE of 0.132 mg/kg, and an RMSE of 0.198 mg/kg. The RF model also exhibited the greatest R^2 value (0.772). The CART model predictions closely followed, with ME, MAE, RMSE, and R^2 values of 0.013 mg/kg, 0.154 mg/kg, 0.230 mg/kg and 0.644, respectively. The three prediction maps generally exhibited similar and realistic spatial patterns of soil Cd contamination. The heavily Cd-affected areas were primarily located in the alluvial valley plain of the Fuchun River and its tributaries because of the dramatic industrialization and urbanization processes that have occurred there. The most important variable for explaining high levels of soil Cd accumulation was the presence of metal smelting industries. The good performance of the RF model was attributable to its

ability to handle the non-linear and hierarchical relationships between soil Cd and environmental variables. These results confirm that the RF approach is promising for the prediction and spatial distribution mapping of soil Cd at the regional scale.

Introduction

Cadmium (Cd) is a toxic metal element that causes extensive concern because of its extremely harmful effects on animals and humans [1]. Due to the low permissible exposure limit of Cd, overexposure may occur even in situations in which trace quantities of Cd are found and can result in metal fume fever, chemical pneumonitis, pulmonary edema, and death [2]. Cd accumulation through the food chain is also harmful to animal and human health. Especially in southern China and northeastern Vietnam, the problem of human exposure to Cd via rice (*Oryza sativa*) intake is of increasing concern [3,4]. The natural concentration of Cd in soils is relatively low; Cd comprises only approximately 0.1 mg/kg of the Earth's crust, and its concentration mainly depends on the geochemistry of the parent material [5,6]. Consequently, soil Cd contamination primarily results from a variety of human activities. Mining [7], smelting [8], electroplating, and scrap metal recycling are coincident with the most important sources of environmental pollution by metals and metalloids [9,10]. Therefore, the challenge is to understand the spatial distributions and high local variabilities in soil Cd concentrations that are caused by the influences of human activities. Such understanding is needed for the preparatory work for remediation. Thus, the evaluation and application of deterministic environmental factors to model the spatial distributions of soil properties (including soil Cd) has been proposed to be an efficient methodology that can serve as an alternative solution to expensive and waste of time soil sampling [11–14].

Numerous statistical techniques for estimating the spatial distributions of soil properties at different scales have been developed and tested within a digital soil-mapping (DSM) framework [15]. In these techniques, linear regression is one of the most frequently used model because of its simplicity, efficiency, and straightforward interpretation [16]. Linear regression models assume that the relationships between the predictor variables and response variables are linear. However, the relationships between soil properties and environmental parameters are often complex and non-linear due to the influences of many factors, such as climate, parent material, topography, and human activities [17]. Machine-learning techniques, such as classification and regression tree (CART) analysis, have been proposed to overcome the shortcomings of linear regression models and to account for the non-linear relationships between soil properties and environmental parameters. CART models can use a wide range of data types and improve the prediction accuracies of spatial models [18,19]. Compared with CART, random forest (RF) modeling, which was developed from CART, is more robust, more resistant to overfitting, and less sensitive to noise in the data [20].

All of these predictive models require an understanding of the factors that control the distributions of the predicted soil properties. Several studies have focused on land use and Cd accumulation in the soil [21–25] because land use data are readily available and extensive and generally represent the impacts of human activities on the soil environment. However, due to lack of detailed description of human activities on the land, land use variables do not always have the ability to play a role in soil Cd prediction [26,27]. When land use variables are employed, significant differences in soil Cd concentrations according to different land use types, for example, woodlands, paddy fields, orchards, vegetable fields, and industrial areas, are

expected. However, the impacts of human activities are too complex to be represented by such coarse land use categories, particularly in the rapidly developing area of eastern China. More detailed information regarding land use, including industrial components, factory distributions, transportation, and urbanization, should be incorporated to predict the spatial distributions of soil Cd concentrations [17].

In the present paper, we tested three empirically based models, i.e., a stepwise linear regression (SLR) traditional linear regression model and two machine learning tools, CART and RF, in the prediction and mapping of the spatial distribution of the soil Cd concentration in Fuyang County. The objectives of this study were to develop and compare the performances of the three models in the estimation of the soil Cd content using factors that likely influence soil Cd concentrations and to identify the likely sources of this contamination.

Materials and Methods

Ethics statement

All research involved in this study complied with the laws of the People's Republic of China, and permission for the field experiments in Fuyang County was obtained from the Agricultural Bureau of Fuyang County. The study site did not involve endangered or protected species.

Site description

The study site was located in Fuyang County in northern Zhejiang Province, which is one of the most economically developed provinces in China (Fig 1). This county (119°25′00″-120° 19′30″ E, 29°44′45″-30°11′58.5″ N) has an area of 1,831 km^2 and a landscape characterized by a mountain and valley topography. The elevation ranges from 1.6 to 1,063.4 m. In recent decades, urbanization and industrialization have occurred at an unprecedented pace in Fuyang County. In the study area, paper mills, metal smelters, hardware machinery factories, and building material factories have been extensively developed. Moreover, the total number of motor vehicles has rapidly increased, and heavy traffic exists throughout the study area [28].

Data collection

Soil samples (n = 276) were collected from the agricultural land across the study area in 2005 based on the uniformity of plot distributions and the land use types in the study area (Fig 1). Each of the samples were collected at a depth of 0–20 cm from five sampling points within 5 m around a specific sampling location and then mixed. A global positioning system (GPS) was used to precisely locate every sampling location. All samples were air dried at room temperature. Stones and plant residuals in the soil samples were manually removed, and the soil samples were then ground to pass through a 2 mm sieve. These samples were analyzed for soil organic matter (SOM), pH, and Cd. The soil pH was determined with a pH meter with a soil/water ratio of 1:2.5, and the SOM was determined using the $K_2Cr_2O_7$-H_2SO_4 oxidation method. Total Cd was determined by digesting the soil sample with a mixture of nitric acid (HNO_3) and perchloric acid ($HClO_4$) followed by measurements, which were determined by inductively coupled plasma mass spectrometry (Agilent 7500a, USA) [29]. The accuracy of determinations was checked using national standard product (GBW 07401). The quality control gave good precision (S.D. $<$ 10%) for all samples. Detection limit for total soil Cd concentrations was 0.02 mg/kg.

A digital elevation model (DEM) with a 25 m spatial resolution and land use data including information on the land use types (Fig 2a), industry types (Fig 2b), and town center and main

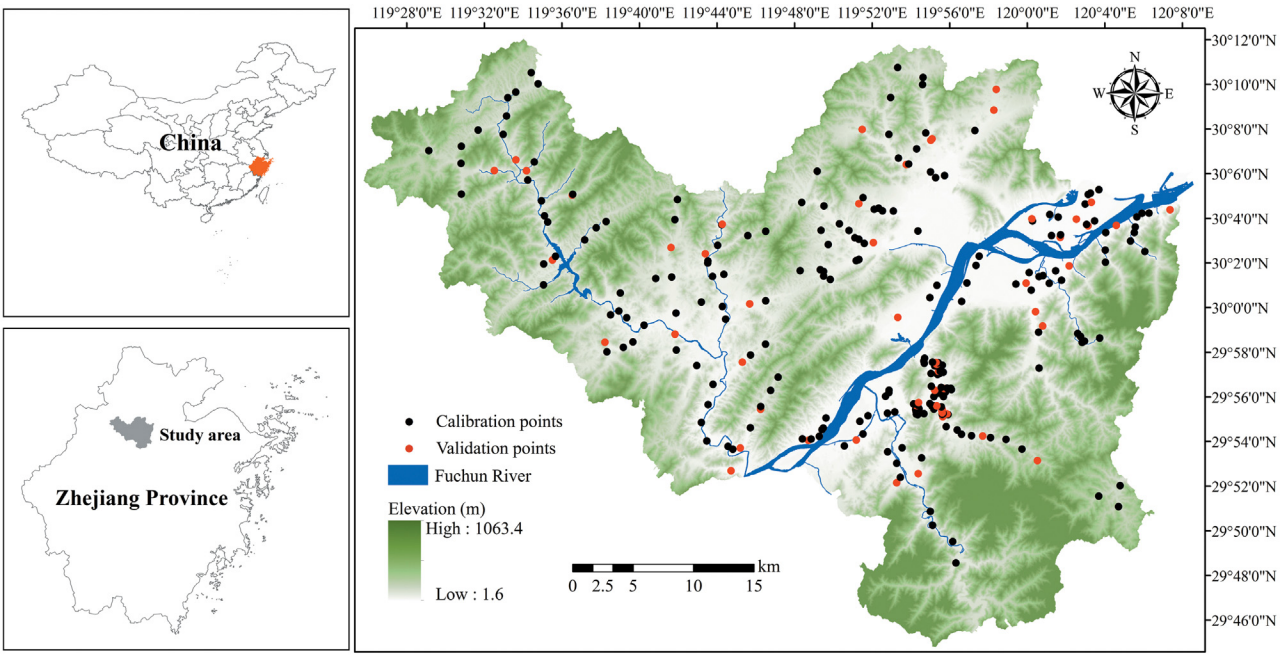

Fig 1. Study area location and sampling points.

doi:10.1371/journal.pone.0151131.g001

highway (Fig 2c) for the year 2005 were obtained from the Bureau of Land and Resources of Fuyang County.

Model construction

The sampling locations were added to an ArcGIS 10.0 (ESRI Inc., Redlands, CA, USA) geodatabase in which the soil properties were associated with the sampling points. The point location layer was intersected with the land use data and the DEM to obtain the independent variables, and the soil Cd concentrations were the dependent variables. Based on a review of the literature [14,17,30–34], we selected a priori 16 environmental variables (Table 1) which represent soil properties, topographic, and anthropogenic activities to explain the spatial variability in the soil Cd concentration.

Pioneering research on soil Cu, Zn, Pb, and Cd contamination in Fuyang County by Zhang et al. (2008 and 2009) [17,28] demonstrated that Fuyang's soil was elevated in Cu, Zn, Pb, and Cd in the areas where industrial plants, towns, and roadways were concentrated. Thus, in our study, the distances to different kinds of industry plants, highways, and town centers were used as important predictors of soil Cd concentrations. Besides, it is reported that application of manure, fertilizers, pesticides, and herbicides were closely linked with soil Cd accumulation [17], therefore agricultural land uses were also selected in analysis. As a result, five main land use categories were classified in the study area. These were vegetable field, paddy field, dryland, forests and orchard. Again, soil properties and topography are two key factors determining natural contents and transport processes of soil Cd. Nevertheless, the measured Cd concentrations were ranged from 0.006 to 2.216 mg/kg with the mean values of 0.322 ± 0.394 mg/kg in the study area. 68.8% of the measured values exceed their according background values of 0.14 mg/kg for Cd in soil at Zhejiang Province and 30.1% of the samples are higher than the Chinese Environmental Quality Secondary Standard for Soils of 0.3 mg/kg (GB 15618–1995) [24]. Highly elevated Cd concentrations coupled with its high spatial variability suggest that

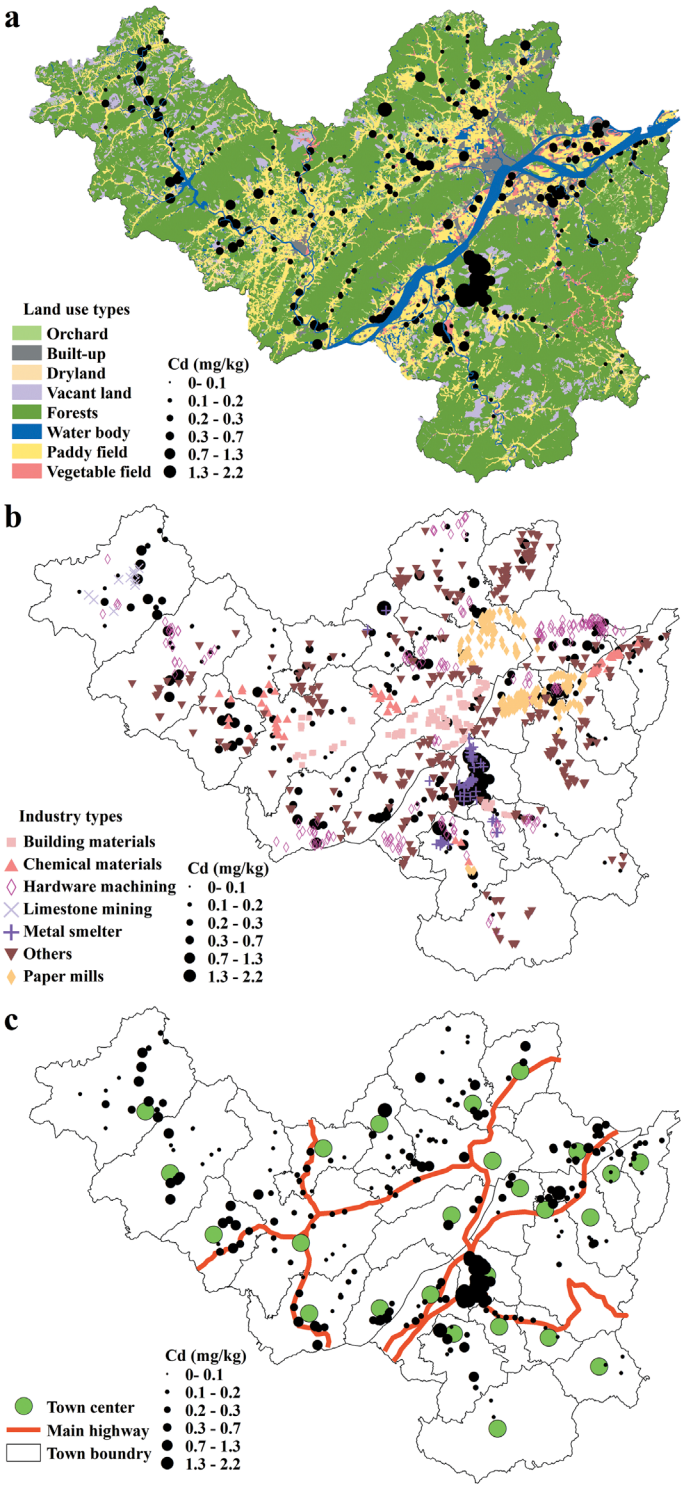

Fig 2. Spatial distribution of the soil Cd concentrations in relation to the (a) land use types, (b) industry types and (c) town center and main highway.

doi:10.1371/journal.pone.0151131.g002

Table 1. The environmental variables selected for model calibration.

Environmental variable	Abbreviation	Unit	Type	Mean	Minimum	Maximum
(a) Soil data						
pH	pH	-	Continuous	6.06	4.40	8.31
Soil organic matter	SOM	%	Continuous	3.12	0.57	6.50
(b) Elevation	ELE	m	Continuous	37.66	4.43	146.26
(c) Land use types						
Vegetable field [a]	Lu_Veg	-	Binary	-	-	-
Paddy field [a]	Lu_Pad	-	Binary	-	-	-
Dryland [a]	Lu_Dry	-	Binary	-	-	-
Forests [a]	Lu_For	-	Binary	-	-	-
Orchard [a]	Lu_Orc	-	Binary	-	-	-
(d) Distance to industry						
Metal smelter	Dmetal	km	Continuous	9.07	0.05	30.69
Hardware machining	Dhardware	km	Continuous	3.39	0.04	11.38
Building materials	Dbuild	km	Continuous	7.43	0.03	26.91
Chemical materials	Dchemical	km	Continuous	6.24	0.06	18.50
Paper mills	Dpaper	km	Continuous	10.73	0.04	40.58
Other industry	Dother	km	Continuous	34.27	0.34	59.43
(e) Distance to highway	Droad	km	Continuous	3.10	0.002	21.97
(f) Distance to town center	Dtown	km	Continuous	3.14	0.33	8.81

[a] Binary variable (0 for absence and 1 for presence).

doi:10.1371/journal.pone.0151131.t001

anthropogenic inputs may be the primary source of Cd in the study area. This has indicated that anthropogenic factors are more likely to predict soil Cd concentrations than natural factors. Thus, we only selected pH, SOM, and elevation in analysis as representative factors of natural environment. Moreover, soil pH and organic matter influence soil Cd concentrations because they are strongly correlated with the solubility and mobility of soil Cd. And elevation is usually the most readily available among topographic variables. In this study, soil pH were ranged from 4.40 to 8.31, with the mean value of 6.06; SOM ranged from 0.57 to 6.50%, with the mean value of 3.12%; and elevation ranged from 4.43 to 146.26 m, with the mean value of 37.66 m (data in S1 Dataset).

Stepwise linear regression. Following a method developed by Montgomery [35], we constructed a SLR model that predicted soil Cd concentrations in the study area with environmental features. Because the original data failed the Kolmogorov-Smirnov normality test ($P < 0.05$), they were log transformed to ensure a normal distribution. Then, the construction of SLR model was conducted in SPSS® (Version 16.0) software.

Following the removal of the variables that were not statistically significant, the SLR parameter estimates were conducted again. ELE, pH, Dmetal, and Dtown were retained in the model. To apply the estimate parameters across the study area, raster data layers for all of the predictor variables (i.e., ELE, pH, Dmetal, and Dtown) were created. Within ArcGIS, the soil pH distribution was interpolated by ordinary kriging, which is a commonly used interpolation method. Raster datasets were created for Dmetal and Dtown by the Euclidean Distance tool in ArcGIS. To match the spatial resolution of the DEM used in the study, a 25 m^2 cell size was used in the construction of all raster layers. Additionally, only areas described as agricultural land were retained in the dataset. Areas described as built-up lands and water bodies were removed from the dataset due to no soil was present there.

Based on the parameters that resulted from the SLR, the raster calculator tool in ArcGIS was used to predict the distribution of soil Cd concentrations. The resulting raster dataset reflected the spatially distribution of log-transformed Cd concentration values. Finally, the resulting raster dataset was reclassified in ArcGIS such that the cells reflected a binary dataset of values that exceeded or below the Chinese soil Cd guide limit of 0.3 mg/kg (Fig 3a). The construction of the binary dataset was conducted for two reasons: (1) to facilitate the follow-up management and (2) for comparison with the CART model, which separates the Cd values into classes.

Classification and regression tree. Classification and regression tree (CART) is a machine-learning algorithm that can be used to split a complex decision into several branched and simplified decisions and may lead to an easier solution [36]. The trees are grown by recursively partitioning a dataset of the dependent variable into a series of binary subsets. Compared to SLR and other traditional general linear models, one potential advantage of CART is the ability to discover unexpected and fresh patterns in non-normal and complex data.

In the CART model, the Cd concentration for each sample was categorized as either low (0–0.3 mg/kg) or high (> 0.3 mg/kg) and used as a target variable, and the 16 environmental variables were utilized as model inputs. The CART analysis was conducted with the R statistical software using the rpart package [37]. Next, the prediction rules calculated by the CART model were translated into a series of branched and simplified statements that were constructed using the raster calculator tool in ArcGIS. The resultant map revealed the areas in which the soil Cd was predicted to exceed the Chinese soil Cd guide limit of 0.3 mg/kg (Fig 3b).

Random forests. Developed from CART analysis, which produces a single tree, random forests (RF) combine a forest of uncorrelated trees created with the CART procedure [38]. Each tree is constructed by a randomly selected subset of training data. The remaining training data, which are called "out-of-bag", are used to estimate prediction error and variable importance [39]. Three training parameters, i.e., (i) the number of trees to grow (n_{tree}), (ii) the number of predictor variables used to split each node (m_{try}), and (iii) the minimum number of observations at the terminal nodes of the trees (*nodesize*), were set to 1,000, 12, and 5, respectively.

We used the randomForest package in the R environment to create an RF model based on the 222-sample training dataset [20]. Next, the model was applied to a continuous ASCII dataset that contained the same independent variables used to construct the RF model. The output result from the randomForest package was an ASCII format file and then was converted to a raster dataset using ArcGIS. Finally, the raster dataset was reclassified to a binary map that also displayed the areas in which the soil Cd was predicted to exceed the Chinese soil Cd guide limit of 0.3 mg/kg (Fig 3c).

Validation

To evaluate the model performances, 54 sampling points (Fig 1) were randomly selected from the original 276 soil samples as the validation samples using the subset features tool of the Geostatistical Analyst extension in ArcGIS. The accuracy assessment was based on analysis of the error matrix, which was a square array of dimensions $n \times n$ (n was the number of classes). This matrix revealed the relationship between the estimated and measured Cd concentrations. The total accuracy and the kappa coefficient were selected to evaluate the prediction accuracy [17]. The former is the ratio of the total number of correctly predicted Cd concentrations to the total number of validation samples (n = 54), and the latter uses all of the information in the error matrix and ranges from 0 to 1. A value of 1 implies perfect agreement, and values below 1 imply less than perfect agreements. Fleiss [40] characterized kappa coefficients below 0.40 as poor, 0.40 to 0.75 as fair to good, and over 0.75 as excellent. Additionally, the mean error (ME),

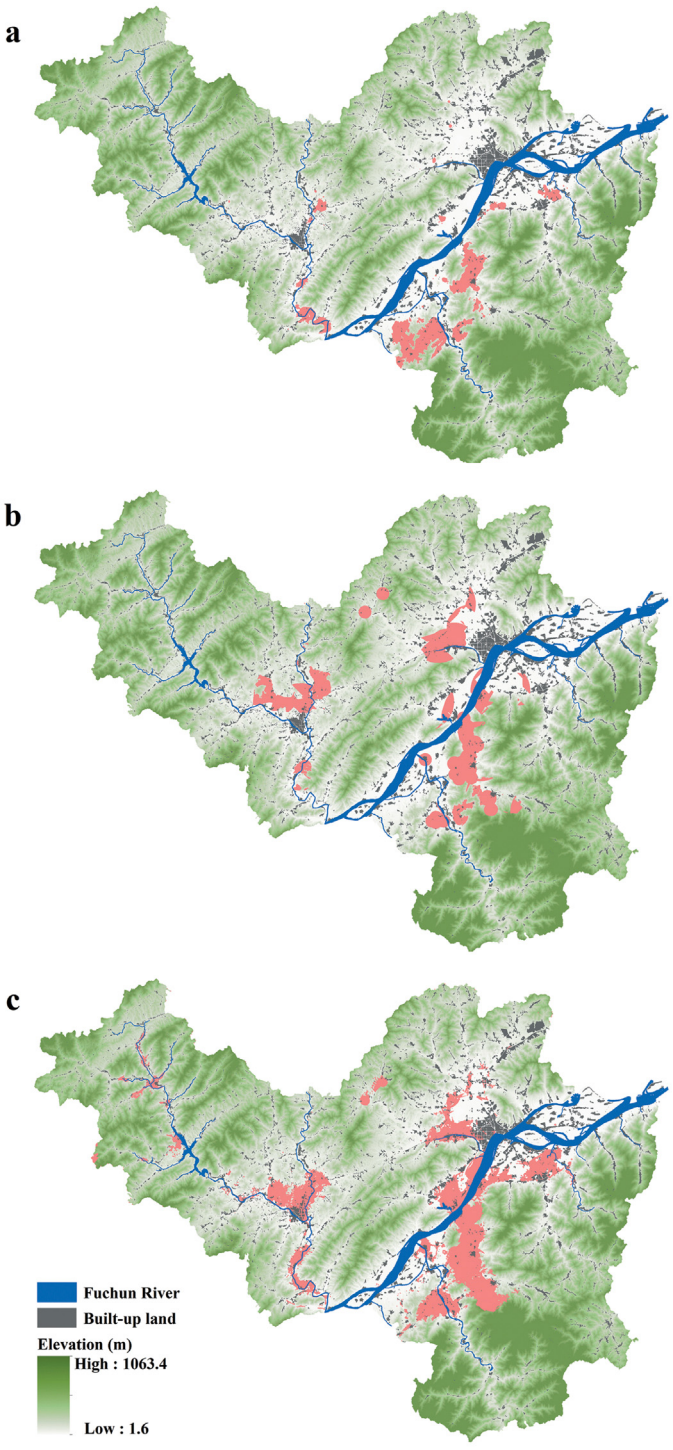

Fig 3. (a) SLR, (b) CART, and (c) RF predictions of Cd in agricultural soils in Fuyang County. The areas in red are predicted to exceed the Chinese soil Cd guide limit of 0.3 mg/kg.

doi:10.1371/journal.pone.0151131.g003

root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) were calculated to assess the accuracy of the predicted Cd concentrations [11,41].

Results and Discussion

Relative importance of the predictor variables

The estimated ELE, pH, and Dmetal parameters were all highly significant at the $P < 0.001$ level, and Dtown was significant at $P < 0.05$ (Table 2). The constant of the model was also significant at the $P < 0.001$ level. These results indicated that ELE, pH, Dmetal, and Dtown substantially influenced the spatial distribution of Cd concentrations.

The simulated tree model contained 8 terminal nodes and 7 independent variables, including Dmetal, pH, SOM, ELE, Dbuild, Dtown, and Droad. Dmetal was the most important factor for classifying the statuses of Cd contamination (Fig 4). Samples located less than 0.7 km from any metal smelter that influenced soil Cd concentration were predicted to contain Cd concentrations that exceeded the Chinese soil Cd guide limit of 0.3 mg/kg.

The variations in the misclassification error and the numbers of terminal nodes as the predictors were excluded one by one from the constructed CART models and are listed in Table 3. The numbers of terminal nodes for the 7 tree models were within the range of 7 to 9, which indicates that these similarly complicated tree models did not grow too tall or overly complex. Based on the magnitude of the increase in the misclassification error, Dmetal was the most important variable for explaining the spatial variations in the Cd concentrations. This finding was unsurprising for statistical and theoretical reasons. Statistically, the Pearson correlation identified between the Cd concentrations and Dmetal values was significant ($P < 0.01$), which indicated that Dmetal was a good predictor of Cd concentration. Theoretically, this high correlation can be ascribed to the strong relationship between soil Cd accumulation and metal smelting industries. It is reported that nonferrous metal industry is strongly correlated with soil Cd contamination throughout the world [42–45], and such metal smelting is one of the pillar industries in Fuyang County. Large quantities of Cd were released into the surrounding environment due to intense emissions of smelting gases and improper disposal of solid wastes [46]. The contributions of the remaining predictors to the models were more or less the same because their exclusions marginally increased the misclassification error. These findings indicate that although these variables influenced soil Cd accumulation, they did not exhibit significant effects on Cd accumulation compared with Dmetal.

It was difficult to judge the importance of each predictor in the RF model because RF algorithms do not reveal the functional relationships between the target and predictor variables. Due to this limitation in interpretability, RF models are called "black box" approaches [41].

Spatial prediction and Cd concentration mapping

The predictions of the Cd levels in the agricultural soils based on the SLR, CART, and RF models are displayed in Fig 3. Generally, the three prediction maps were similar and realistic in

Table 2. Estimated parameters of the SLR model.

Variable	Parameter	Std. error	t value	p-value
Constant	-3.619	0.379	-9.541	0.000
ELE	-0.007	0.002	-4.314	0.000
pH	0.473	0.061	7.79	0.000
Dmetal	-0.038	0.008	-5.012	0.000
Dtown	-0.087	0.035	-2.503	0.013

doi:10.1371/journal.pone.0151131.t002

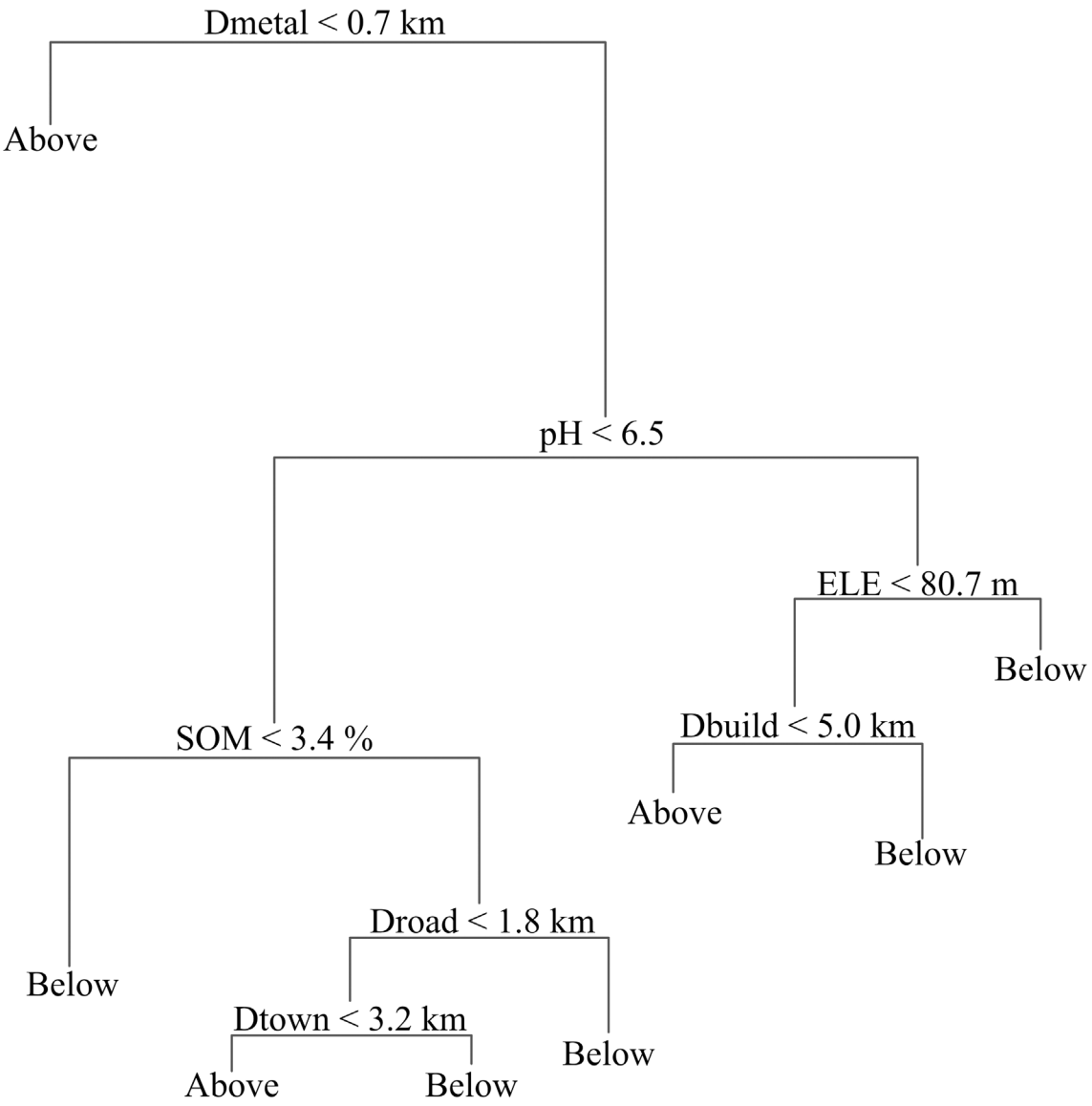

Fig 4. CART model developed to predict Cd in agricultural soils in Fuyang County. The lengths of the lines or "branches" are proportional to the variance explained, and longer branches explain more variance. Below: 0–0.3 mg/kg; Above: > 0.3 mg/kg.

doi:10.1371/journal.pone.0151131.g004

terms of the spatial patterns of Cd contamination. The areas in which the soil Cd concentration exceeded 0.3 mg/kg were primarily located in the alluvial valley plain of Fuchun River and its tributaries. This finding clearly reflected the effects of industrial operations, especially metal smelting activities, on soil Cd accumulation because these areas were industrially well developed according to the statistical data (Fig 2b). Moreover, the most dramatic urban sprawl occurred in these regions (Fig 2c). Rapid urban expansion also explained the high Cd concentrations [33] in these areas. In contrast, the areas in which the soil Cd concentration was below 0.3 mg/kg were distributed in the higher-altitude areas, and this pattern could be partly explained by presence of indigenous forests that are hundreds of years old in these areas.

The area predicted to exceed the Chinese soil Cd guide limit of 0.3 mg/kg varied among the models. The SLR model predicted the smallest area (32.69 km^2 or 1.8% of the total modeled

Table 3. The relative importance of the variables for explaining soil Cd variation as indicated by variations in misclassification errors and the numbers of terminal nodes in the CART model following the exclusion of predictors.

Variable	Misclassification error rate	Number of terminal nodes
Missing Dbuild	8.11%	9
Missing ELE	9.01%	9
Missing Dtown	9.91%	8
Missing SOM	10.36%	7
Missing Droad	10.36%	8
Missing pH	10.81%	7
Missing Dmetal	13.51%	8

doi:10.1371/journal.pone.0151131.t003

area) that would exhibit a soil Cd concentration above 0.3 mg/kg, whereas the RF model predicted the greatest area (89.30 km^2 or 4.9%), and the CART model predicted a medium area (68.39 km^2 or 3.8%) that would exceed 0.3 mg/kg.

The CART model predicted a contaminated area of more than twice the size of that predicted by the SLR. This difference may have resulted from the application of a rule in the CART model that states that the soil Cd will be above 0.3 mg/kg within 0.7 km of any metal smelter regardless of the other environmental factors. This rule is consistent with the observations from the sampled data in that elevated levels of Cd were found next to metal smelters (Fig 2b) and also with other studies that have documented serious risks of Cu pollution within 1,500 m of metal smelters [17]. The RF model predicted an even larger area of contamination than the CART model, which may be attributable to the RF model's advantage in handling complex data relationships [11]. In the present study, the relationships between soil Cd accumulation and the influences of human activities were nonlinear and hierarchical and were revealed by the SLR and CART models. In contrast to the SLR model, the RF model required no assumptions regarding the relationships between soil Cd concentration and influencing factors and handled the nonlinear and hierarchical relationships without such assumptions.

Performances of the three models

The validation dataset (n = 54) was used to test the performances of three models. For the SLR, 39 of the 54 validation samples were correctly classified, which resulted in a total accuracy of 72.22% (Table 4). The kappa coefficient, which represents the inter-rater agreement for qualitative (categorical) items, was 0.4048, which indicated a fair to good prediction accuracy from the SLR. Regarding the CART, 38 of the 54 validation samples were correctly classified, which resulted in a total accuracy of 70.37% (Table 4). The kappa coefficient was 0.3949, which indicated a very close to good prediction accuracy of the CART. The RF model correctly classified 41 of the 54 validation samples, which resulted in an overall accuracy of 75.93% (Table 4). The kappa coefficient was 0.5050, which indicated the prediction accuracy of the RF model was the greatest among the three models.

Fig 5 compares the observed and predicted soil Cd concentrations using the validation dataset. This figure also displays the prediction error indices derived from the independent validations of the soil Cd concentrations using the validation dataset. Positive ME values indicate that the models underestimated the Cd concentration. Specifically, the SLR model exhibited the greatest tendency for underestimation, with an ME of 0.074 mg/kg, whereas the RF model exhibited the lowest tendency for underestimation, with an ME of 0.002 mg/kg (Fig 5). Statistically, the MAE is a quantity used to measure the differences between predictions and eventual outcomes. The SLR model exhibited the largest predicted deviation, with an MAE of 0.160 mg/

Table 4. Error matrices for the SLR, CART and RF predictions of the soil Cd concentrations.

	SLR		CART		RF	
	Low	High	Low	High	Low	High
Low	27	5	23	9	25	7
High	10	12	7	15	6	16
Accuracy (%)	72.97	70.59	76.67	62.50	80.65	69.57

SLR: total accuracy: 72.22% and kappa coefficient: 0.4048. CART: total accuracy: 70.37% and kappa coefficient: 0.3949. RF: 75.93% and kappa coefficient: 0.5050. Low: 0–0.3 mg/kg; High: > 0.3 mg/kg

doi:10.1371/journal.pone.0151131.t004

kg, whereas the predictions of the RF model (MAE 0.132 mg/kg) were closest to observed values (Fig 5). Moreover, the RF model exhibited the lowest RMSE (0.198 mg/kg) and the highest R^2 value (0.772). Compared with other studies [11,41,47], the R^2 value of the RF model in the present study was slightly higher. This difference may have resulted from differences in the study areas, topographies, sampling strategies, or quantities and qualities of the utilized environmental variables. Hence, the RF method was optimal for predicting the Cd concentrations of the unvisited locations in this context, followed closely by the CART model, which produced predictions with ME, MAE, RMSE, and R^2 values of 0.013 mg/kg, 0.154 mg/kg, 0.230 mg/kg and 0.644, respectively. The SLR model produced the poorest prediction results, as indicated by the highest values for these three error indices (ME = 0.074 mg/kg, MAE = 0.160 mg/kg, and RMSE = 0.274 mg/kg) and the lowest R^2 (0.542) in the three models (Fig 5).

The models described above are most likely applicable to other county-scale regions, such as counties in southeastern China with similar urbanization and industrialization processes, although the accuracies of the models may depend on regional characteristics. The SLR model provides a convenient and reasonable method when data sources are limited. The RF model predicted the largest contaminated area. It may be less likely to exclude possible Cd contamination and more protective of environment. However, when working with environmental managing departments, the CART is the preferred method due both to its high accuracy and a series of statements that predict pollution classes which are convenient for translation into public policy.

Conclusions

In the present study, the RF method was found to be the best method for spatially predicting and mapping soil Cd concentration patterns in Fuyang County, China. The accuracy of the RF model was satisfactory, and the prediction map produced by the RF method revealed a realistic spatial pattern of soil Cd contamination. Compared to the SLR, the RF model performed much better in predicting and mapping the spatial distribution of soil Cd because the RF model proficiently handled the nonlinear and hierarchical relationships between the soil Cd and the main influencing factors. This result confirmed the reliability of the use of RF to model and predict the spatial distribution of soil Cd using environmental variables. This approach could be selected as an alternative methodology to reduce the cost of intensive soil sampling.

Furthermore, analysis of the importance of each variable identified the presence of metal smelting industries as the most important variable for explaining high soil Cd accumulation in the study area. Intense emissions of smelting gases and improper disposal of solid wastes were most likely responsible for the high Cd concentrations in the soils. The requirement of remediation approaches such as phytostabilization and phytoextraction by hyperaccumulator plants

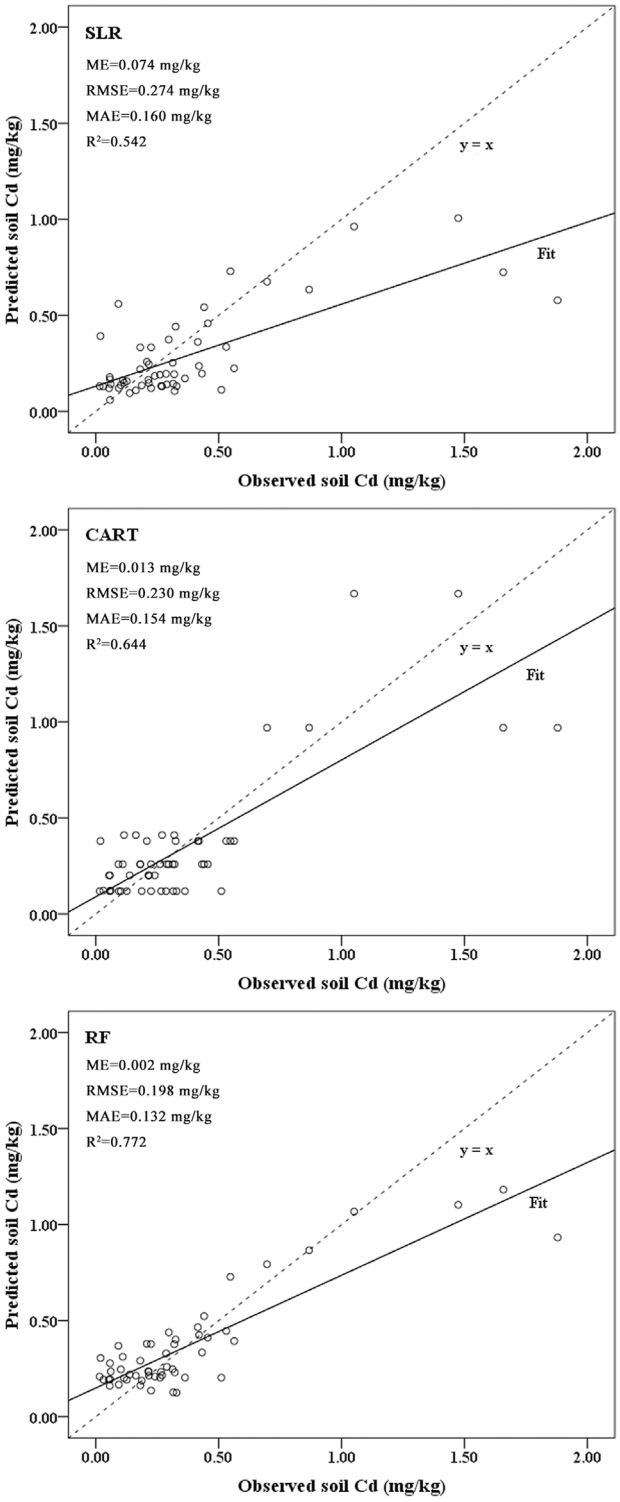

Fig 5. Performances of the SLR, CART, and RF models in the prediction of soil Cd concentrations.

doi:10.1371/journal.pone.0151131.g005

practices in local areas threatened by Cd [48,49]. Our findings could provide necessary information for policy makers in land use management and effectively prevent further Cd pollution.

Supporting Information

S1 Dataset. Original data of model construction.
(XLS)

Acknowledgments

The source data set is provided by (1) Data Center for Resources and Environmental Sciences, Chinese Academy of Sciences (RESDC) (http://www.resdc.cn), (2) Geospatial Data Cloud, Computer Network Information Center, Chinese Academy of Sciences (http://www.gscloud.cn), and (3) Bureau of Land and Resources of Fuyang County.

Author Contributions

Conceived and designed the experiments: LFQ. Performed the experiments: LFQ Kai Wang. Analyzed the data: LFQ Kai Wang. Contributed reagents/materials/analysis tools: WLL Ke Wang. Wrote the paper: LFQ WH GSA. Obtained permission for use of original data: Ke Wang.

References

1. Kirkham MB. Cadmium in plants on polluted soils: effects of soil factors, hyperaccumulation, and amendments. Geoderma. 2006; 137: 19–32.
2. Hayes AW. Principles and methods of toxicology. Philadelphia: CRC Press; 2007.
3. Li ZW, Li LQ, Chen GX, Pan J. Bioavailability of Cd in a soil—rice system in China: soil type versus genotype effects. Plant Soil. 2005; 271: 165–173.
4. Martinez RE, Marquez JE, Hòa HTB, Gieré R. Open-pit coal-mining effects on rice paddy soil composition and metal bioavailability to Oryza sativa L. plants in Cam Pha, northeastern Vietnam. Environ Sci Pollut Res. 2013; 20: 7686–7698.
5. Hans Wedepohl K. The composition of the continental crust. Geochim Cosmochim Acta. 1995; 59: 1217–1232.
6. McLennan SM. Relationships between the trace element composition of sedimentary rocks and upper continental crust. Geochem Geophy Geosy. 2001; 2, 109.
7. Pourret O, Lange B, Bonhoure J, Colinet G, Decrée S, Mahy G, et al. Assessment of soil metal distribution and environmental impact of mining in Katanga (Democratic Republic of Congo). Appl Geochem. 2016; 64: 43–55.
8. Ettler V. Soil contamination near non-ferrous metal smelters: A review. Appl Geochem. 2016; 64: 56–74.
9. Nriagu JO. A history of global metal pollution. Science. 1996; 272: 223–224.
10. Alloway BJ. Heavy metals in soils. 2th ed. London: Blackie Academic and Professional Press; 1995.
11. Guo PT, Li MF, Luo W, Tang QF, Liu ZW, Lin ZM. Digital mapping of soil organic matter for rubber plantation at regional scale: an application of random forest plus residuals kriging approach. Geoderma. 2015; 237–238: 49–59.
12. Liénard A, Brostaux Y, Colinet G. Soil contamination near a former Zn-Pb oretreatment plant: Evaluation of deterministic factors and spatial structures at the landscape scale. J Geochem Explor. 2014; 147: 107–116.
13. Bou Kheir R, Shomar B, Greve MB, Greve MH. On the quantitative relationships between environmental parameters and heavy metals pollution in Mediterranean soils using GIS regression-trees: the case study of Lebanon. J Geochem Explor. 2014; 147: 250–259.
14. Bou Kheir R, Greve MH, Abdallah C, Dalgaard T. Spatial soil zinc content distribution from terrain parameters: a GIS-based decision-tree model in Lebanon. Environ Pollut. 2010; 158: 520–528. doi: 10.1016/j.envpol.2009.08.009 PMID: 19773104

15. McBratney AB, Mendonça Santos ML, Minasny B. On digital soil mapping. Geoderma. 2003; 117: 3–52.
16. Thompson JA, Pena-Yewtukhiw EM, Grove JH. Soil—landscape modeling across a physiographic region: topographic patterns and model transportability. Geoderma. 2006; 133: 57–70.
17. Zhang XY, Lin FF, Jiang YG, Wang K, Wong MTF. Assessing soil Cu content and anthropogenic influences using decision tree analysis. Environ Pollut. 2008; 156: 1260–1267. doi: 10.1016/j.envpol.2008.03.009 PMID: 18455844
18. De'ath G, Fabricius KE. Classification and regression trees: A powerful yet simple technique for ecological data analysis. Ecol. 2000; 81: 3178–3192.
19. Drake JM, Randin C, Guisan A. Modelling ecological niches with support vector machines. J Appl Ecol. 2006; 43: 424–432.
20. Breiman L, Cutler A. Breiman and Cutler's random forests for classification and regression. R package version 46–7; 2013. Available: https://cran.r-project.org/web/packages/randomForest/.
21. Xia XH, Chen X, Liu RM, Liu H. Heavy metals in urban soils with various types of land use in Beijing, China. J Hazard Mater. 2011; 186: 2043–2050. doi: 10.1016/j.jhazmat.2010.12.104 PMID: 21242029
22. Zhao YC, Wang ZG, Sun WX, Huang B, Shi XZ, Ji JF. Spatial interrelations and multi-scale sources of soil heavy metal variability in a typical urban—rural transition area in Yangtze River Delta region of China. Geoderma. 2010; 156: 216–227.
23. Qishlaqi A, Moore F, Forghani G. Characterization of metal pollution in soils under two landuse patterns in the Angouran region, NW Iran: a study based on multivariate data analysis. J Hazard Mater. 2009; 172: 374–384. doi: 10.1016/j.jhazmat.2009.07.024 PMID: 19647938
24. Zhang XY, Lin FF, Jiang YG, Wang K, Feng XL. Variability of total and available copper concentrations in relation to land use and soil properties in Yangtze River Delta of China. Environ Monit Assess. 2009; 155: 205–213. doi: 10.1007/s10661-008-0429-9 PMID: 18618282
25. Kelly J, Thornton I, Simpson PR. Urban geochemistry: A study of the influence of anthropogenic activity on the heavy metal content of soils in traditionally industrial and non-industrial areas of Britain. Appl Geochem. 1996; 11: 363–370.
26. Schwarz K, Pickett STA, Lathrop RG, Weathers KC, Pouyat RV, Cadenasso ML. The effects of the urban built environment on the spatial distribution of lead in residential soils. Environ Pollut. 2012; 163: 32–39. doi: 10.1016/j.envpol.2011.12.003 PMID: 22325428
27. Pouyat R, Belt K, Pataki D, Groffman P, Hom J, Babd L. Urban land-use change effects on biogeochemical cycles. Terrestrial ecosystems in a changing world. The IGBP series. Berlin: Springer Verlag; 2007.
28. Zhang XY, Lin FF, Wong MTF, Feng XL, Wang K. Identification of soil heavy metal sources from anthropogenic activities and pollution assessment of Fuyang County, China. Environ Monit Assess. 2009; 154: 439–449. doi: 10.1007/s10661-008-0410-7 PMID: 18597177
29. Chen T, Liu XM, Li X, Zhao KL, Zhang JB, Xu JM, et al. Heavy metal sources identification and sampling uncertainty analysis in a field-scale vegetable soil of Hangzhou, China. Environ Pollut. 2009; 157: 1003–1010. doi: 10.1016/j.envpol.2008.10.011 PMID: 19026475
30. Quenea K, Lamy I, Winterton P, Bermond A, Dumat C. Interactions between metals and soil organic matter in various particle size fractions of soil contaminated with waste water. Geoderma. 2009; 149: 217–223.
31. Qiu LF, Gan MY, Wang K, Deng JS, Hong Y, Xu JF, et al. Source identification of soil Cu, Zn, Pb, and Cd from anthropogenic activities by decisicion tree analysis in Fuyang County, China. Fresenius Environ Bull. 2012; 21: 1390–1398.
32. Micó C, Recatalá L, Peris M, Sánchez J. Assessing heavy metal sources in agricultural soils of an European Mediterranean area by multivariate analysis. Chemosphere. 2006; 65: 863–872. PMID: 16635506
33. Chen HY, Teng YG, Lu SJ, Wang YY, Wang JS. Contamination features and health risk of soil heavy metals in China. Sci Total Environ. 2015; 512–513: 143–153. doi: 10.1016/j.scitotenv.2015.01.025 PMID: 25617996
34. Cai LM, Xu ZC, Bao P, He M, Dou L, Chen LG, et al. Multivariate and geostatistical analyses of the spatial distribution and source of arsenic and heavy metals in the agricultural soils in Shunde, Southeast China. J Geochem Explor. 2015; 148: 189–195.
35. Montgomery DC, Peck EA, Vining GG. Introduction to linear regression analysis, 5th ed. Hoboken, NJ: John Wiley & Sons; 2012.
36. Breiman L, Friedman JH, Olshen RA, Stone CG. Classification and regression trees. Belmont, CA: Wadsworth International Group Press; 1984.

37. Therneau T, Atkinson B, Ripley B. Recursive partitioning and regression trees. R package Version 41–8. Available at: http://cran.stat.ucla.edu/web/packages/rpart/index.html; 2014.

38. Breiman L. Random forests. Mach Learn. 2001; 45: 5–32.

39. Cutler DR, Edwards TC Jr, Beard KH, Cutler A, Hess KT, Gibson J, et al. Random forests for classification in ecology. Ecology. 2007; 88: 2783–2792. PMID: 18051647

40. Fleiss JL. Statistical methods for rates and proportions. 2nd ed. New York: John Wiley & Sons; 1981.

41. Were K, Bui DT, Dick ØB, Singh BR. 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. 2015; 52: 394–403.

42. Wei CY, Wang C, Yang LS. Characterizing spatial distribution and sources of heavy metals in the soils from mining-smelting activities in Shuikoushan, Hunan Province, China. J Environ Sci. 2009; 21: 1230–1236.

43. Sterckeman T, Douay F, Proix N, Fourrier H. Vertical distribution of Cd, Pb and Zn in soils near smelters in the north of France. Environ Pollut. 2000; 107: 377–389. PMID: 15092984

44. Bacon JR, Dinev NS. Isotopic characterisation of lead in contaminated soils from the vicinity of a non-ferrous metal smelter near Plovdiv, Bulgaria. Environ Pollut. 2005; 134: 247–255. PMID: 15589652

45. Barcan V. Nature and origin of multicomponent aerial emissions of the copper—nickel smelter complex. Environ Int. 2002; 28: 451–456. PMID: 12503910

46. Bi XY, Feng XB, Yang YG, Qiu GL, Li GH, Li FL, et al. Environmental contamination of heavy metals from zinc smelting areas in Hezhang County, western Guizhou, China. Environ Int. 2006; 32: 883–890. PMID: 16806473

47. Schwarz K, Weathers KC, Pickett STA, Lathrop RG, Pouyat RV, Cadenasso ML. A comparison of three empirically based, spatially explicit predictive models of residential soil Pb concentrations in Baltimore, Maryland, USA: understanding the variability within cities. Environ Geochem Health. 2013; 35: 495–510. doi: 10.1007/s10653-013-9510-6 PMID: 23775390

48. Houben D, Couder E, Sonnet P. Leachability of cadmium, lead, and zinc in a longterm spontaneously revegetated slag heap: Implications for phytostabilization. J Soils Sediments. 2013; 13: 543–554.

49. Wang K, Zhu ZQ, Huang HG, Li TQ, He ZL, Yang XE, et al. Interactive effects of Cd and PAHs on contaminants removal from co-contaminated soil planted with hyperaccumulator plant Sedum alfredii. J Soils Sediments. 2012; 12: 556–564.