Environ Monit Assess (2022) 194: 518
https://doi.org/10.1007/s10661-022-10125-5

Hyperspectral sensing of heavy metals in soil by integrating AI and UAV technology

Ho Wen Chen · Chien-Yuan Chen · Kieu Lan Phuong Nguyen · Bin-Jiun Chen · Chang-Hsuan Tsai

Received: 1 June 2021 / Accepted: 17 May 2022 / Published online: 22 June 2022

Abstract Given the limitation of conventional soil pollution monitoring through mapping which is a costly, time-consuming work, the study aims to establish an image recognition model to identify the source of pollution automatically. The study choses a contaminated land and then use a non-destructive instrument that can quickly and effectively measure the content of heavy metals. A two concentration prediction models of Ni, Cu, Zn, Cr, Pb, As, Cd, and Hg using hyperspectral imaging were developed, Decision Tree and Back Propagation Neural Network, in combination of particle swarm optimization employed for optimization algorithm. As a result, random forest is more accurate than the forecast result of back propagation neural network. This study has established an excellent Cu and Cr model, which can accurately capture the pollution source. In addition, through aerial photographs, we also found that there were also high pollution reactions on the banks of the river. The developed model is beneficial for high pollution areas which can be quickly found, thereby following investigation and remediation work can be carried out with less time and cost consuming comparing with the conventional soil monitoring.

Keywords Hyperspectral sensing · UAVs · Heavy metals · Soil pollution · Neural network · Random forest

H. W. Chen · B.-J. Chen · C.-H. Tsai
Department of Environmental Science and Engineering, Tung-Hai University, Taichung, Taiwan

H. W. Chen
Center for Smart Sustainable Circular Economy, Tung-Hai University, Taichung, Taiwan

C.-Y. Chen
Department of Civil and Water Resources Engineering, National Chiayi University, Chiayi, Taiwan

K. L. P. Nguyen (✉)
Faculty of Environmental and Food Engineering, Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam
e-mail: nklphuong@ntt.edu.vn

Introduction

Due to rapid economic and population growth, soil pollution caused by anthropogenic activities has become one of the most critical environmental issues needed to address globally (Rodríguez Eugenio et al., 2018; Sutherland, 2000; Buaisha et al., 2020). Soil pollution refers to the presence of hazardous chemical substances or the concentration of a chemical substance in the soil is higher than normal. This can result from natural geogenic sources or from anthropogenic sources. Natural and geogenic pollution sources can be defined as the presence of soil pollution caused by natural events, such as forest fire, volcanic eruption, and weathering of heavy metal-containing materials and ores (Albanese et al., 2007; Díez et al., 2009; Hafez & Awad, 2016). On the other hand, anthropogenic pollution sources

refer to the emission and accumulation of chemicals and toxic elements in the soil from human activities. Compared to natural soil pollution, anthropogenic soil pollution is regarded as an urgent environmental issue in developing countries. According to Myint (1973), developing countries characterize themselves as vibrant economies, where economic activities and population are growing rapidly; their economies are highly dependent on primary and secondary sectors that are often related to pollution and environmental degradation. In this regard, many works of literature have demonstrated that anthropogenic activities, such as industrial activities, mining, agricultural and livestock activities, transport infrastructure, and the generation or disposal of waste and sewage, are found highly associated with soil pollution and land contamination in developing countries (Bundschuh et al., 2012; Hu et al., 2013). In addition, due to the lack of public awareness of environmental protection and social demand for practicing environmental policies, environmental problems have not been alleviated successfully in most developing countries (Bell & Russell, 2002). Eventually, these environmental problems will backfire on the economic growth itself. For example, fertile farmlands and non-contaminated soil and water are often seen as key to secure high agricultural productivity, abundant food supply, and human health, whereas polluted ones mean the opposite (Chowdhury et al., 2016; Rajaganapa et al., 2011; Zwolak et al., 2019). Previous and current literature has also indicated that soil pollution is an urgent global issue for human health and food security concerns, particularly in developing countries (Antoniadis et al., 2019; Lai et al., 2010; Lal, 2000; Qin et al., 2021). For example, Lin et al. (2002) conducted a landscape analysis to detect the distribution of soil pollution in central Taiwan, a major food production area in Taiwan, and found that contaminated farmlands are usually located near industrial and urban areas; Rajaganapa et al. (2011) discovered that heavy metal contamination had polluted soil and water in many areas in India, affecting agricultural and livestock products and posing health risks to the public. Hence, to address food security concerns and health issues related to soil pollution, identifying polluted lands and locating contaminated sites will be essential. However, soil pollution often cannot be directly observed with the naked eye, which increases the difficulty in detecting polluted areas and addressing this issue.

Mapping is one of the ways for soil pollution monitoring. Conventional soil mapping is to physically gather soil samples in a grid pattern and then to send the soil to a laboratory for analyzing chemical properties. Jia et al. (2021) indicated that contaminant concentration prediction using geostatistical interpolation methods (GIMs), e.g., Kriging interpolation, for unsampled points are costly, time-consuming work and not well suitable to highly heterogeneous soil environments.

With the increase of world population and the need for food consumption, the use of UAV technology in agriculture has become increasingly necessary. There are many applications that UAV could help improve crop yield, among of which could be named: (i) soil and field analysis: data collected by UAV could help in early soil analysis as well as in planning seed planting patterns, in making irrigation plan, in determining quantity of fertilizers and/or pesticides needed; (ii) planting: with the ability to cover large areas of land, UAVs have simplified crop planting and reduced planting costs cause they have reduced man powers for the work; (iii) crop and spot spraying: again due to large covering ability of UAVs, crop spraying has been simplified, with the aim of sensor, UAVs improve the spraying accuracy and conserves resources; (iv) crop monitoring: UAVs with thermal imaging cameras and/or multiple spectral cameras enable the farmer to monitor his farm. The farmer can check the state of crops in the farm, as well as areas that need urgent attention; (v) irrigation: UAVs equipped with thermal imaging cameras could help farmers determine areas with low soil moisture, find crops that are hydrated, locate areas that are waterlogged; and (vi) health assessment: UAV-carried properly devices can detect variations (related to fungal and bacterial diseases) and associate these variations to farmers for early interventions, UAVs also offer new and modern methods of accurately monitoring and assessing pest damage.

UAVs equipped with spectroscopy technology will generate a large number of data sets and increase weather, hydrology, geography, and other data according to mission requirements. Therefore, effective and high-speed algorithms are required to analyze big data. Artificial

intelligence algorithms are suitable for solving such problems. For crop yield and biomass, image recognition is used to detect whether crops are suffering from diseases, and classification techniques are used to distinguish healthy and diseased crops (Lan et al., 2020). Qun'ou et al. (2021) employed twelve machine learning algorithms, i.e., support vector machine (SVM), artificial neural networks (ANN), Bayesian ridge regression (BRR), Lasso regression (Lasso), elastic net (EN), linear regression (LR), decision tree regression (DTR), K neighbors regression (KNR), random forest regression (RFR), extra trees regression (ETR), AdaBoost regression (ABR), and gradient boosting regression (GBR) to explore the most accurate retrieval model of the total nitrogen concentration among them which used the UAV hyper spectral remote sensing and ground monitoring data.

In the current study, we aim to establish an image recognition model to identify the source of pollution automatically. The study chose a contaminated land and then use a non-destructive instrument that can quickly and effectively measure the content of heavy metals. A two concentration prediction models of Ni, Cu, Zn, Cr, Pb, As, Cd, and Hg using hyperspectral imaging were developed decision tree and back propagation neural network, while particle swarm optimization was employed for optimization algorithm.

The novel contribution of this study is to establish image recognition model which can accurately capture the pollution source automatically, thereby following investigation and remediation work can be carried out with less time and cost consuming comparing with conventional soil mapping monitoring. However, the difficulty encountered in this study is that the spectroscopy technique can only detect surface soil, and this technology cannot simulate the dynamic changes of heavy metal components in surface soil. In addition, this study does not consider the soil composition for the time being.

Methodology

Study area

According to Taiwan's Environmental Protection Agency statistics in 2018, there are currently more than 7000 regulated soil and groundwater pollution sites in Taiwan. The approximate location is shown in Fig. 1 on the left side. In the early days, Taiwan was an agricultural society, in which most of its agriculture was based on irrigation. With the industrialization of society, small factories gradually entered rural areas. In the absence of a sound sewage treatment system design that separates irrigation and drainage, industrial wastewater often causes large-scale farmlands to be polluted, among which contaminated farmlands in Changhua County and Taoyuan City are the most serious. Therefore, this study selects the contaminated land in Dacun Township, Changhua County, in which is regulated by the Soil and Groundwater Remediation Fund Management Board of the Environmental Protection Agency in Taiwan, as shown in Fig. 1 on the right side. The area is about 2.06 hectares, including multiple striped, rectangular, irregular-shaped lands. The land is announced to be polluted by heavy metals such as copper (Cu) and nickel (Ni), which may enter the food chain through various means and will seriously increase the risk of cancer (Briffa et al., 2020).

Multispectral imaging with aerial photography

The camera used in this study is the Sequioa camera produced by Parrot with four wave bands includes green: 550, red: 660, red edge: 735, nir: 790 (nm). Its imaging principle is that solar radiation will be absorbed and reflected when it hits a substance. The reflected light will be received by a sensor, then resolved into different wavelength bands by a beam splitter, finally producing a two-dimensional image.

Figure 2 is an aerial image with 3507×2480 pixels, and the shooting size is 750×1250 m. After software analysis, the layers representing different wavebands can be obtained. Because the combination of reflectance will reflect each grid's physical and chemical characteristics, the concentration of heavy metal in a pixel will be figured by the rules embedded in the reflectance combination of different wavebands.

Pollution concentration prediction model using hyperspectral imaging

Feature engineering

This study adopts 254 locations of samples data, which was obtained on-site, from Taiwan's Environmental

Protection Agency statistics for prediction modeling. To execute the concentration prediction model, the study collected the spectral data of the studied locations and its heavy metals concentrations measured including Ni, Cu, Zn, Cr, Pb, As, Cd, and Hg. Finally, these data are divided into a training dataset and a test dataset at a ratio of 4:1. Additionally, in order to improve accuracy of proposed model, some extreme outliers are removed, and normalization method is also used to convert the raw data to a value between 0 and 1with the formula as Eq. (1).

$$z_i = \frac{(x_i - x_{min})}{(x_{max} - x_{min})} \tag{1}$$

Since the range between variables in the data is not identical, and the mean of each range is also different, the data is standardized to improve the comparability of the data and the accuracy of the model. The formula is as follows.

$$x_i = \frac{\left(x_i - x_{mean}\right)}{x_{std}} \tag{2}$$

Fig. 2 Aerial Images. **a** Original image; **b** aerial image with green filter; **c** aerial image with red filter; **d** aerial image with green filter and edge detection; **e** aerial image with red filter and edge detection ▶

Pollution concentration prediction model using hyperspectral imaging

Random forest Random forest is a classifier that contains multiple decision trees, where the output category is determined by the mode of individual trees. The algorithm of each tree is constructed as the following: "N" represents the number of training samples, whereas "M" represents the problem of the variable; therefore, we define the number of variables to be included in decision-making and set a value "m" that is far smaller than "M" to be used on a node; the model will repeatedly sample from N training cases; if sampling N times, a training set will be formed, also known as bootstrapping sampling; finally, the tree will be constructed to predict the remaining category and estimate the error. For each node, the random selection of "m" is based on the variable of this

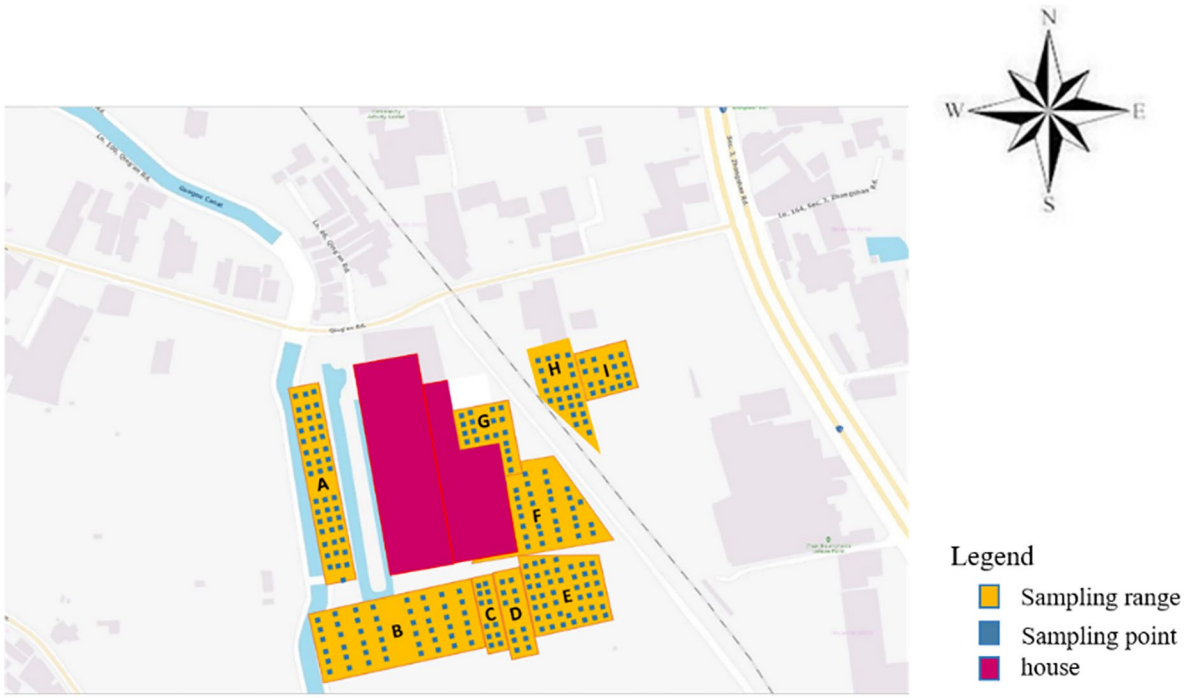


Fig. 1 Schematic diagram of the study area

(a)

(b)

(c)

(d)

(e)

node, which the optimal method of splitting is based on. In the process, each tree will grow intact without pruning. Therefore, the construction of each decision tree in the random forest is based on the following three criteria:

(a) In the process of data training, the training dataset of decision tree is constructed by randomly selecting "n" numbers of data, and the training data can also be used by other trees.
(b) If the data contains "M" observations, "m" observations will be randomly selected (m<<M) as decision variables to build up the decision tree, using this mode as the basis for building a decision tree, finding the best split rule at each node.
(c) Each tree grows to its maximum without any pruning.

There are two main factors in optimizing the target random forest training, that is, the classification ability of each tree and the correlation between any two trees. Intuitively, when each tree has a lower error rate and grows more robustly, the overall classification ability of the random forest will rise. If the correlation between the two trees increases simultaneously, their complementary ability is reduced, leading the error rate to increase. Therefore, the setting of the aforementioned "m" value will strongly affect the robustness of the tree and the correlation between trees. To reduce the value of "m," the available eigenvalues of individual decision trees will be reduced, and the robustness of the tree will also be reduced. At the same time, the allowable difference between trees will also decrease, causing the correlation between each other to reduce. Therefore, the key problem of constructing a random forest is to find the optimal value of "m." The key to solving this problem is mainly to calculate the rate of out-of-bag error. An important advantage of random forest is that there is no need to cross-validate or use an independent test set to obtain an unbiased estimator. It can be evaluated internally, which means that an unbiased estimator can be established during the generation process. When constructing each tree, we used different bootstrapping sampling for the training set. Thus, for each tree (assuming for the *k*th tree), about 1/3 of the training samples did not participate in the generation of the *k*th tree, which are called the "oob" samples of the *k*th tree.

Also, this sampling feature allows us to perform "oob" estimation, which is calculated as follows:

(a) Calculating the classification of oob samples
(b) Using majority voting to determine the classification result of the sample
(c) Calculating the ratio of the number of errors to total number of samples and setting it as the error rate of oob.

However, random forest has another classifier that combines linear regression to return data to a line and make predictions. All parameters, attributes, and interfaces of the regressor are all consistent with the classifier. Only the indicators that measure the quality of the branches are different, and there are three standards supported:

(a) Using mean square error: the difference in the mean squared error between the parent node and the child node will be used as the criterion for feature selection; the L2 loss is minimized by using the mean value of the child node in this way
(b) Using Friedman mean square error: This indicator uses Friedman's improved mean square error for problems in potential branches
(c) Using mean absolute error: this indicator uses the median of the child nodes to minimize the L1 loss

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(f_i - y_i)^2 \tag{3}$$

where N is the number of samples, i is each data sample, f_i is the value returned by the model, and y_i is the actual value label of the sample point i. Therefore, the essence of MSE is the difference between the real data of the sample and the regression result.

In the regression tree, MSE is not only a measurement of branch quality, but also the most commonly used indicator to measure the regression quality of a regression tree. When cross-validation or other methods are used to obtain the results of the regression tree, the mean square error is often selected as the evaluation. Usually the pursuit of linear regression is that the smaller the MSE, the better the model.

However, the score interface of the regression tree returns R-squared, instead of MSE, where R-squared is defined as follows:

$$R^2 = 1 - \frac{u}{v} \tag{4}$$

$$u = \frac{1}{N}\sum_{i=1}^{N}(f_i - y_i)^2 \tag{5}$$

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

where u is the residual sum of squares, v is the total sum of squares, N is the number of samples, i is each data sample, f_i is the value returned by the model, and y_i is the actual value label of the sample point i. R-squared can be positive or negative (if the model's residual sum of squares is much larger than the model's total sum of squares, it means that the model performs badly, and R-squared will be negative). On the other hand, the mean square error is always positive. Among them, random forest is a combination of decision tree prediction models. In order to prevent the random forest model from overfitting, this study sets the maximum number of decision trees to 100 and limits the maximum depth of each subtree to 6 layers. The random forest uses the entropy algorithm as the basis for judging the purity of the root node, and the random forest will preferentially select the node cutting method with high purity.

Back propagation neural network Back propagation neural network (BPNN) belongs to supervised learning, which is composed of three layers of neural units, as shown in Fig. 3. The first layer is an input layer composed of input units, and these input units can receive various features in the sample. The second layer is the hidden layer, and the output calculation formula of the hidden layer is shown as (7). The third layer is the output layer, and its calculation formula is the same as that of the hidden layer. BPNN is divided into two parts: the first part is "forward pass," which is used to calculate the current linear regression error, whereas the second part is to adjust the weight according to the error and correct the error of the network output to the minimum value.

$$H_h = f(net_h) = \frac{1}{1 + e^{-net_h}} \tag{7}$$

The formula of net_h is as (8):

$$net_h = \sum_{i=1}^{N_{nid}} X_i W_{ih} - \theta_h \tag{8}$$

where:

X_i: The output value of the ith neuron in the input layer

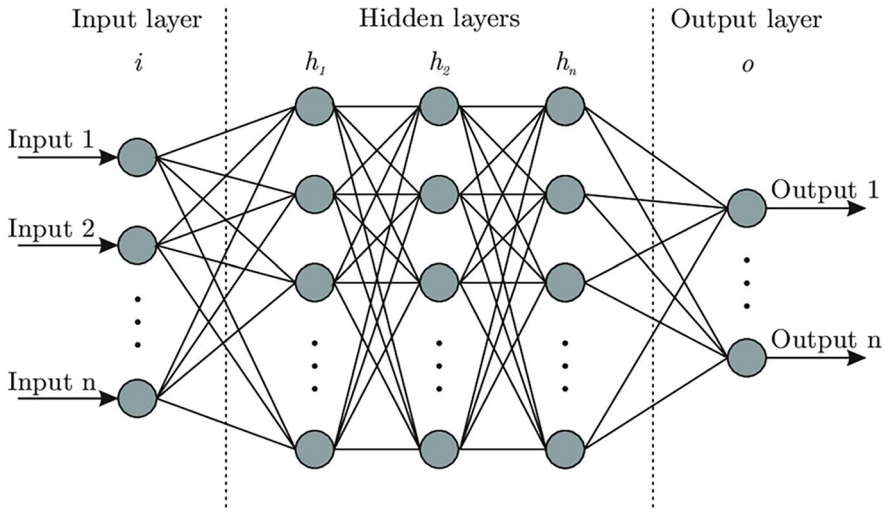


Fig. 3 Schematic graph of BPNN

net_h: Weighted product sum of the hth neuron in the hidden layer

$f(\_)$: Conversion function of the hidden layer

W_{ih}: Weighted function between the ith neuron in the input layer and the hth neuron in the hidden layer

θ_h: Partial weighted value of the hth neuron in the hidden layer

N_{nid}: Number of hidden layer neurons

This study will use the PSO algorithm to optimize the structure of the BPNN neural network. The optimized parameters include the number of hidden layers, hidden layer neurons, and other parameters. This approach aims at selecting the model with the best prediction result.

Hyperparameter optimization-PSO algorithm When building a predictive model, there will be hyperparameters that need to be set, but the range of each parameter is different and may contain multiple parameters, resulting in nearly infinite combinations. Theoretically, the method of exhaustion can be used to find the best solution in the whole domain, but in actual situations, there are not infinite resources that can be consumed. For example, time, performance consumption, etc. are the key points that need to be considered. However, the optimization algorithm can effectively solve this problem. By simulating the learning behavior of organisms, the best approximate solution can be found under limited time and efficiency. The particle swarm optimization (PSO) used in this study is an optimization algorithm developed by Kennedy et al. (1995). At first, it was used to express the flight behavior of the bird flock. By referring to the best solution of personal history and the best solution of the group history, the flight behavior of the flock of birds was expressed, and the inertial weight was added to better control the search behavior.

The conceptual flow of PSO is as follows:

(i) Problem-defining: regarding N parameter optimization problems as an N-dimensional solution space and defining M particles to assist in solving the problem
(ii) Initialization: randomly generating the rate and position of each particle
(iii) Scoring: using the fitness function to evaluate the current score of each particle
(iv) Update: updating the velocity and displacement of particles
(v) First comparison: comparing each particle with its previous optimal position and retain relatively better particles
(vi) Second comparison: comparing all current pbest with gbest of the previous cycle and updating gbest

If the stop condition is met (the number of iterations or the progress rate becomes smaller), the search stops, if not, repeat steps (iii) to (vi).

The speed updating formula is as follows:

$$v = w * v + c1 * rand * \left(pbest - x_{position}\right) + c2 * rand * \left(gbest - x_{position}\right) \quad (9)$$

where:

v: initial velocity
w: inertia weight
c1: speed weight
rand: random numbers between 0 and 1
pbest: best position in personal history
c2: Speed weight
gbest: Best solution in the group history
x_{position}: current position of the particle

The PSO algorithm's parameters were set in the study as follows: w to be 0.7298; c1 and c2 to be 1.496, and the number of iterations to be 1000. Overview of the study procedure is shown in Fig. 4.

Results and discussion

Statistical analysis–resolution of hyperspectral images

As shown in Table 1, there are a total of 204 refined data for modeling. It can be found that the pollution of Cu and Ni are the main pollutants in this study. Although the maximum value of Cu reaches 1163.0 (ug/g) and the maximum value of Ni is as high as 521.0 (ug/g), the maximum value of the two is less than 100 (ug/g). It indicates that the concentration distributions of Cu and Ni in space are uneven. Identifying polluted hot spots and carry out effective remediation is even difficult for engineers, especially

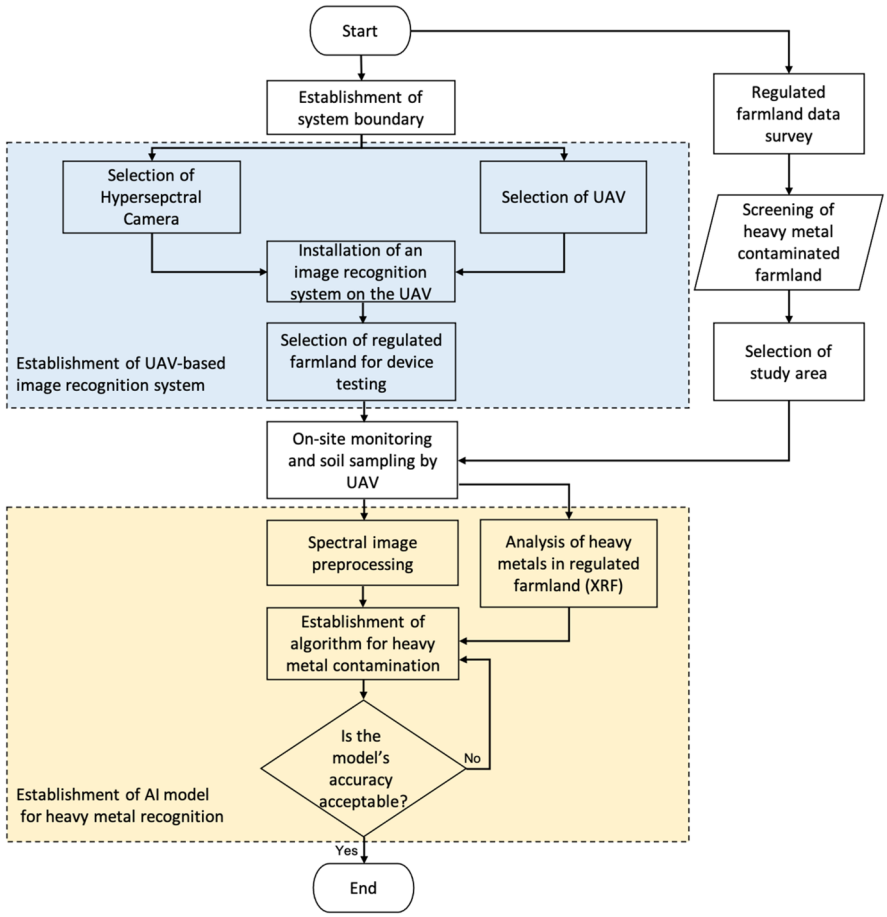


Fig. 4 Overview of the study procedure

for large-scale pollution sites. As shown in Table 2, the water inlet is a highly polluted area. The lowest value of Ni is as high as 144.0 (ug/g), the average value is 281.0 (ug/g), and the highest value is 527.0 (ug/g). Cu pollution is also particularly serious. Although the maximum concentration of Ni and Cu is lower than the concentration in farmland, the ratio of higher than the standard value at the water inlet is significantly higher than that in farmland. However, the concentration of Cd and Hg metals is not detected. Therefore, this study chooses Ni and Cu to establish research model.

Table 1 Concentration distribution of heavy metals in farmland (unit: mg/kg)

	Ni	Cu	Zn	Cr	Pb	As	Cd	Hg
Count	204	204	204	204	204	204	0	0
Mean	174.5	208.8	174.1	157.2	41.9	12.3	0	0
Std	126.2	169.7	29.0	28.4	3.9	6.9	0	0
Min	54.0	46.0	130.0	123.0	14.6	7.7	0	0
25%	72.8	76.0	152.0	138.8	39.5	10.4	0	0
50%	122.5	162.0	164.0	149.0	41.4	12.0	0	0
75%	237.0	268.3	187.3	167.3	44.2	13.1	0	0
Max	521.0	1163.0	284.0	366.0	52.1	107.0	0	0
Standard[1]	130.0	220.0	1000.0	175.0	1000.0	30.0	10.0	10.0

[1]Monitoring standard in soil for various heavy metals

Table 2 Concentration distribution of heavy metals nearby water inlet (unit: mg/kg)

	Ni	Cu	Zn	Cr	Pb	As	Cd	Hg
Count	50	50	50	50	50	50	0	0
Mean	314.4	450.9	208.9	176.8	43.4	11.1	0	0
Std	137.8	215.3	37.3	32.1	3.1	1.6	0	0
Min	144.0	216.0	166.0	131.0	37.5	8.9	0	0
25%	190.0	265.0	177.0	148.8	41.2	10.2	0	0
50%	281.0	349.0	192.0	178.0	43.3	10.8	0	0
75%	448.8	670.0	240.5	189.8	44.8	11.6	0	0
Max	527.0	853.0	278.0	247.0	50.7	16.0	0	0
Standard[1]	130.0	220.0	1000.0	175.0	1000.0	30.0	10.0	10.0

[1]Monitoring standard in soil for various heavy metals

Model prediction

This study records the spectral values of XRF screening positions, and uses the reflectance values of the 550, 660, 735, and 790 spectral bands in the spectral values as the input variables of the model. XRF utilized is a handheld model that determine the.

sample composition by measuring the fluorescent X-ray emitted when it is activated by a primary X-ray source. Afterward, the measured data of XRF is applied for the output variables of the model for modeling, and then establishing prediction models for different heavy metals respectively. This research uses random forest and backward propagation neural network for modeling. Among them, random forest is a combination of decision tree prediction models. In order to prevent the random forest model from overfitting, this study sets the maximum number of decision trees to 100 and limits the maximum depth of each subtree to 6 layers. The backward pass neural network is composed of many neurons, so it is necessary to clearly define the structure of the neural network to avoid over-complex neural network models that will cause overfitting. In this study, the parameters were set as batchsize = 64; learn_rate = 0.01; iteration = 1000.

The results of the two modes are shown in Table 3 and the representation. Overall The model based on random forest is more accurate than BPNN. In terms of accuracy and precision, copper metal prediction performs better than other heavy metals either by random forest or BPNN. From the results of MAPE in Table 3 and Table 4, prediction model based on random forest can be considered as a good prediction model because the value of MAPE is roughly between 10 and 20%. During the training process of the model, the performance of random forest is better than BPNN. It seems to imply that random forest has an advantage in small sample problems. The model based on random forest can effectively predict the concentrations of Cu and Ni, those are the key contaminants in this study. Unexpectedly, random forest can also predict the concentration of Cr, efficiently.

Table 3 Accuracy analysis for random forest

	Training dataset			Test dataset		
	RMSE	MAPE	R^2	RMSE	MAPE	R^2
Ni	6.31	16.00	0.99	150.02	31.93	0.63
Cu	11.19	24.50	0.99	40.02	15.0	0.81
Zn	23.27	18.61	0.88	72.74	31.91	0.50
Cr	12.32	35.28	0.97	46.82	14.16	0.81
Pb	10.15	13.37	0.64	37.02	15.32	0.00
As	5.62	21.2	0.69	42.62	5.04	0.01

RMSE root mean square error, *MAPE* mean absolute percentage error

Table 4 Accuracy analysis for BPNN

	Training			Test		
	RMSE	MAPE	$R.^2$	RMSE	MAPE	$R.^2$
Ni	96.80	170.2	0.43	172.2	201.35	0.54
Cu	25.00	24.25	0.76	53.33	76.68	0.64
Zn	23.69	31.30	0.53	57.10	60.25	0.52
Cr	45.38	48.53	0.57	84.86	114.02	0.68
Pb	21.90	45.46	0.63	48.70	50.83	0.01
As	7.53	17.46	0.68	13.76	30.26	0.00

RMSE root mean square error, *MAPE* mean absolute percentage error

Poor performance of random forest prediction model occurs at low concentration of heavy metals such as Pb and As. This is reasonable because there is no variation in the concentration of these two heavy metals.

The structure of the random forest model of Cu and Ni is shown in Appendix 1. The model test results of various heavy metals are shown in Fig. 5. From the result of Fig. 5, the R^2 value of each random forest model reveals that Cu prediction model R^2 is as high as 0.81, while Ni is only slightly lower, and it is also as high as 0.59. The results imply that the model based on random forest can be used to screen out the height polluted soils with high concentrations of Cu, Ni, and Cr. The R^2 of Zn is about as 0.50, and overall when the Zn concentration is greater than 75 mg/kg, The predicted value of the Zn forecast model will be underestimated.

Sensitivity analysis of proposed prediction model

In order to understand the impact of the input spectral band team's prediction results, this study analyzes the importance and sensitivity of the variables. The results in Fig. 6 show that the 550 spectral band has a great contribution to the prediction of the six metals, and the contribution degree is the second. 660 band, especially for Ni metal. The 550 spectral band is within the blue spectral range, so future prediction models can incorporate more blue spectral bands to try to improve. Figure 7 shows that Ni metal is less sensitive to the spectral band of 660, Cu metal is more sensitive to long wavelengths, and the relatively sensitive bands of Zn are 735, 550, 660, and 790, 735 for these three heavy metals. The high sensitivity indicates that this band is prone to variability in the prediction results. On the whole, heavy metals are more sensitive to the information of the 550 and 735 bands, and the 735 band ranks in the lower part of the forecast contribution. Therefore, it can be inferred that the 735 band may reduce the accuracy of the prediction model and cause the large fluctuations in the accuracy of the prediction model.

Spatial patterns of heavy metals with hyperspectral imaging

The black spots in Fig. 8 are high pollution spots identified after on-site measurements. After using the prediction model to predict the spatial concentration and comparing with these high pollution concentration points, it is found that the model can effectively predict the concentration of heavy metals in highly polluted soil. Figure 8 shows the results of using no-load images and prediction models to estimate the spatial distribution of heavy metal concentrations. Cu, Ni, Zn, and Cr pollution is concentrated in the three regions G, F, and E. The high pollution concentration of Pb metal is concentrated along the river bank, while As pollution does not have any pollution hot spots. This trend is similar to the results of on-site measurements. The difference is that the method proposed in this research is used to predict the concentration of space. It is found that in addition to the obvious high concentration of heavy metals near the water inlet, the polluted hot areas in the agricultural land are also at a glance.

In order to conduct a comprehensive assessment, this study uses Eq. (10) to conduct a heavy metal hazard risk analysis, and the results are shown in Fig. 9. There are four areas with higher pollution risks (marked by dark blue dotted circle), so these areas should be prioritized for remediation in the near future.

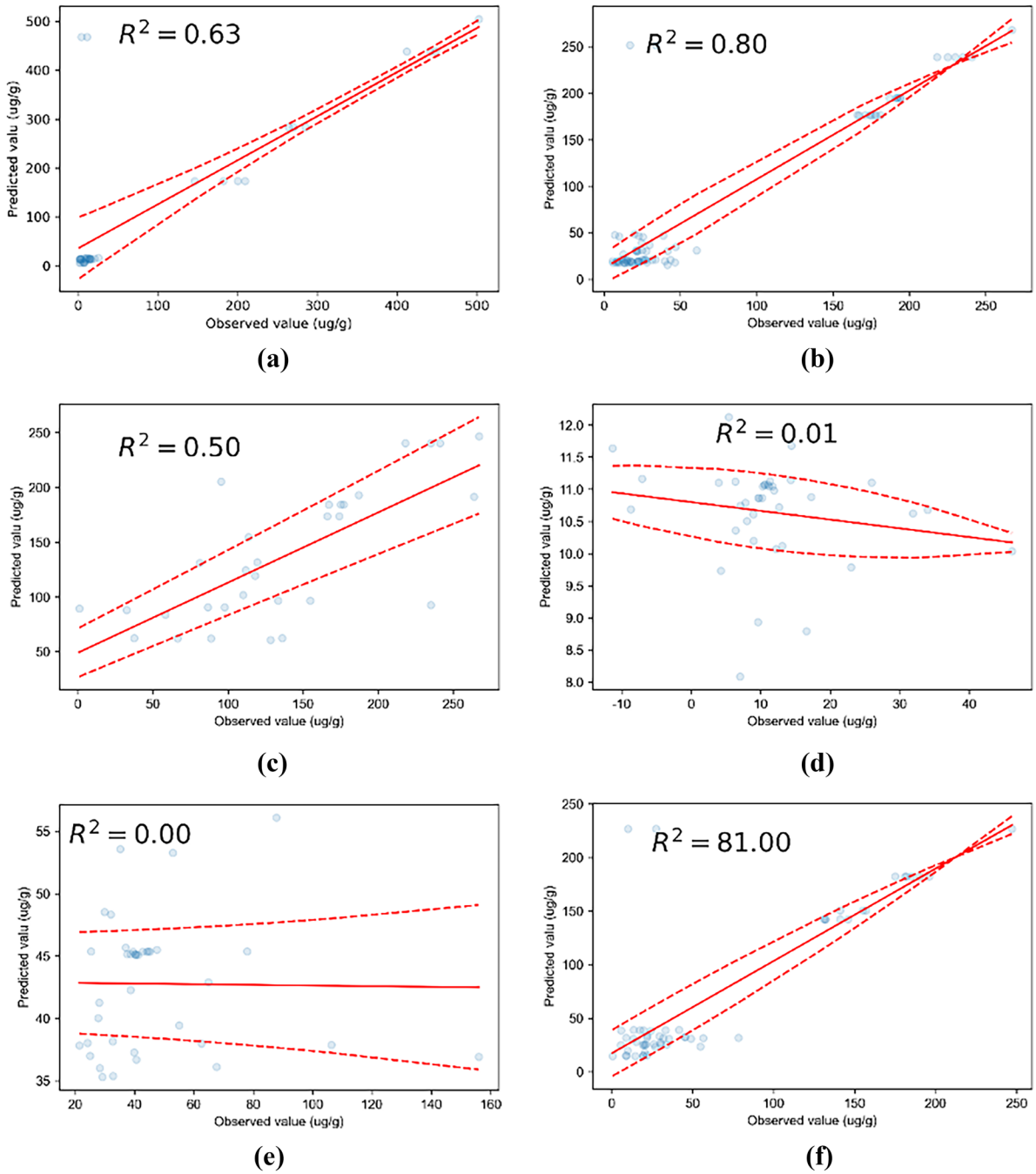


Fig. 5 Error analysis for Random Forest with testing data. **a** Ni; **b** Cu; **c** Zn; **d** As; **e** Pb; **f** Cr

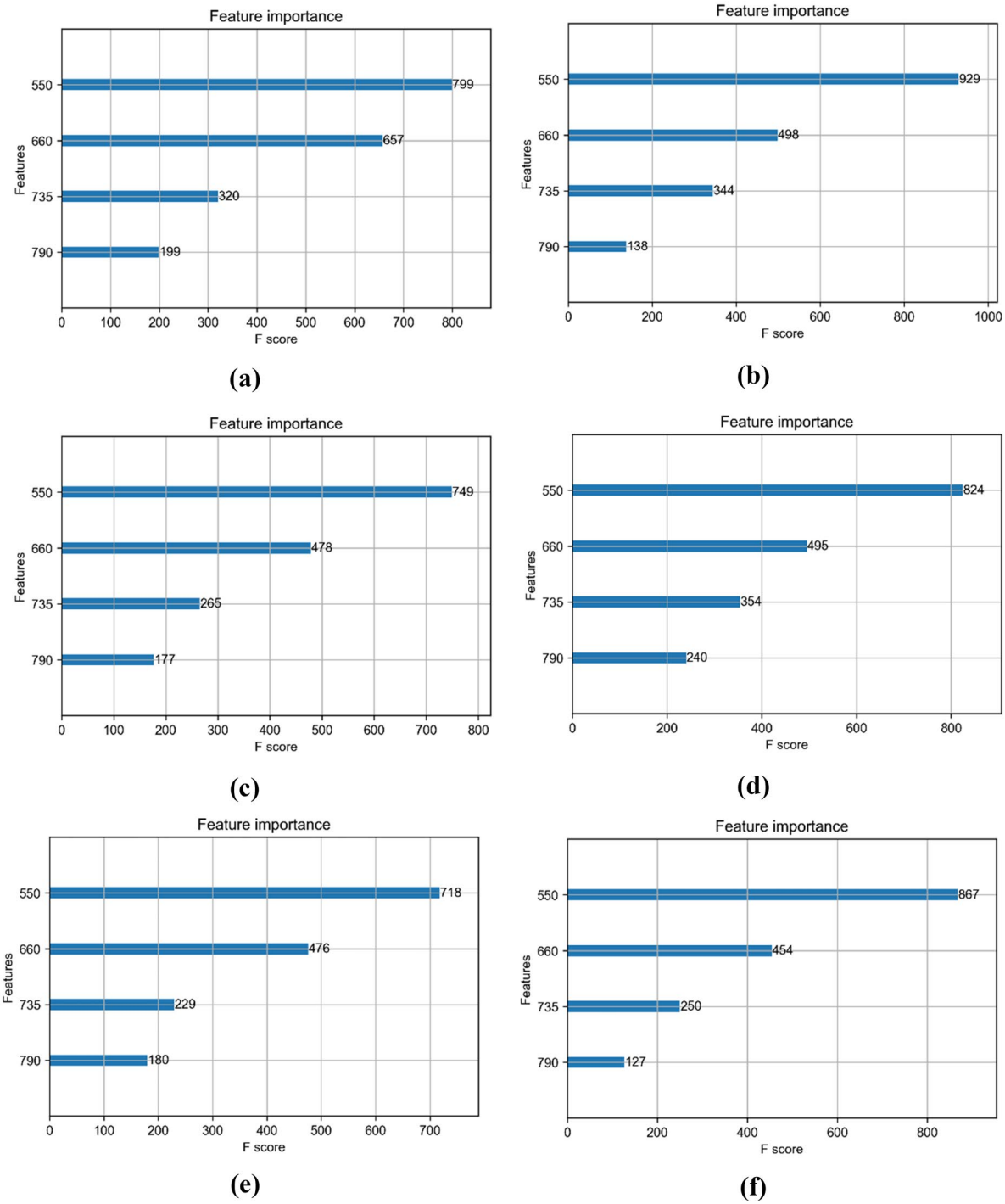


Fig. 6 Importance analysis of input variable for random forest. **a** Ni; **b** Cu; **c** Zn; **d** As; **e** Pb; **f** Cr

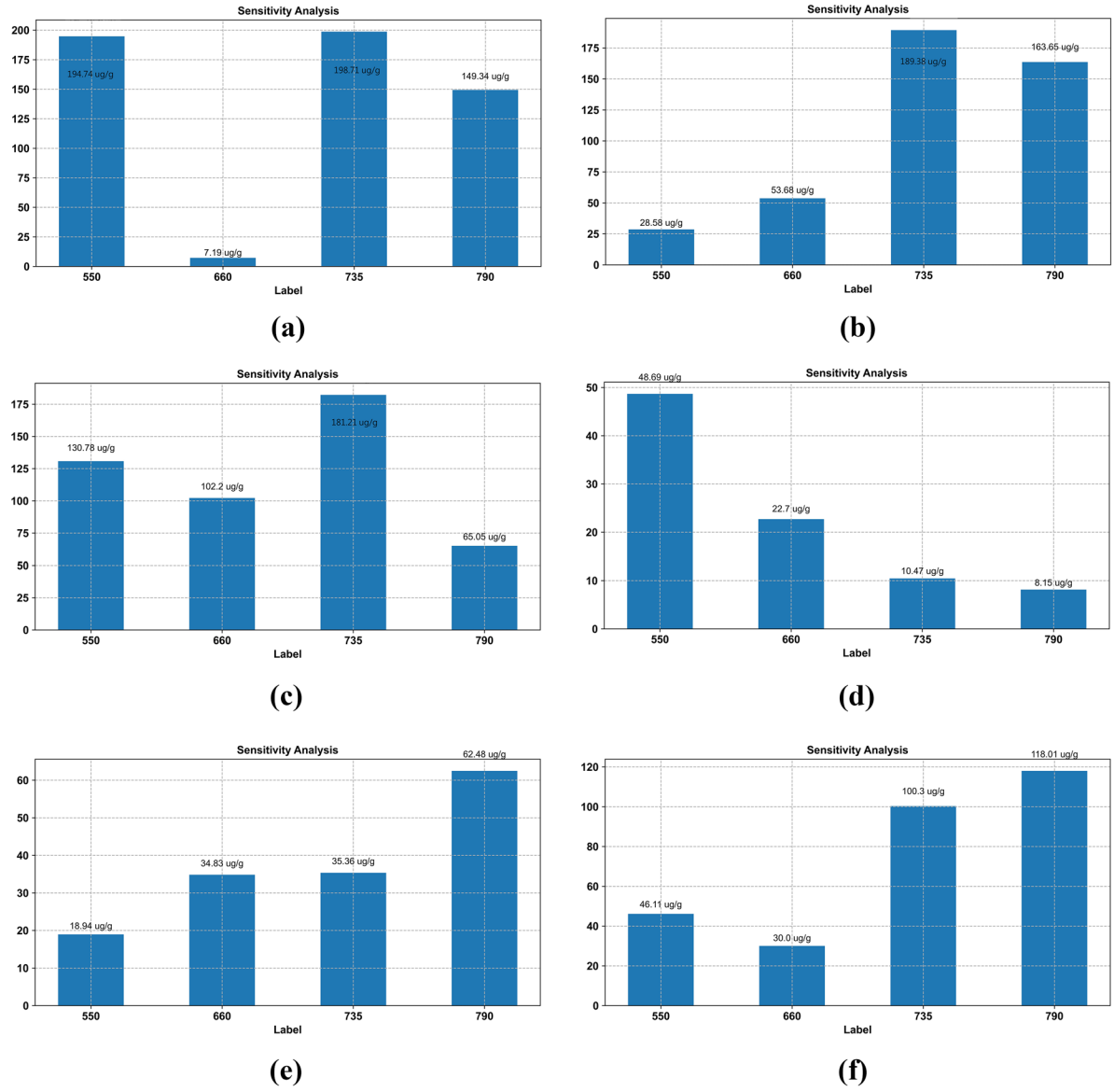


Fig. 7 Sensitivity analysis of input variable for random forest. **a** Ni; **b** Cu; **c** Zn; **d** As; **e** Pb; **f** Cr

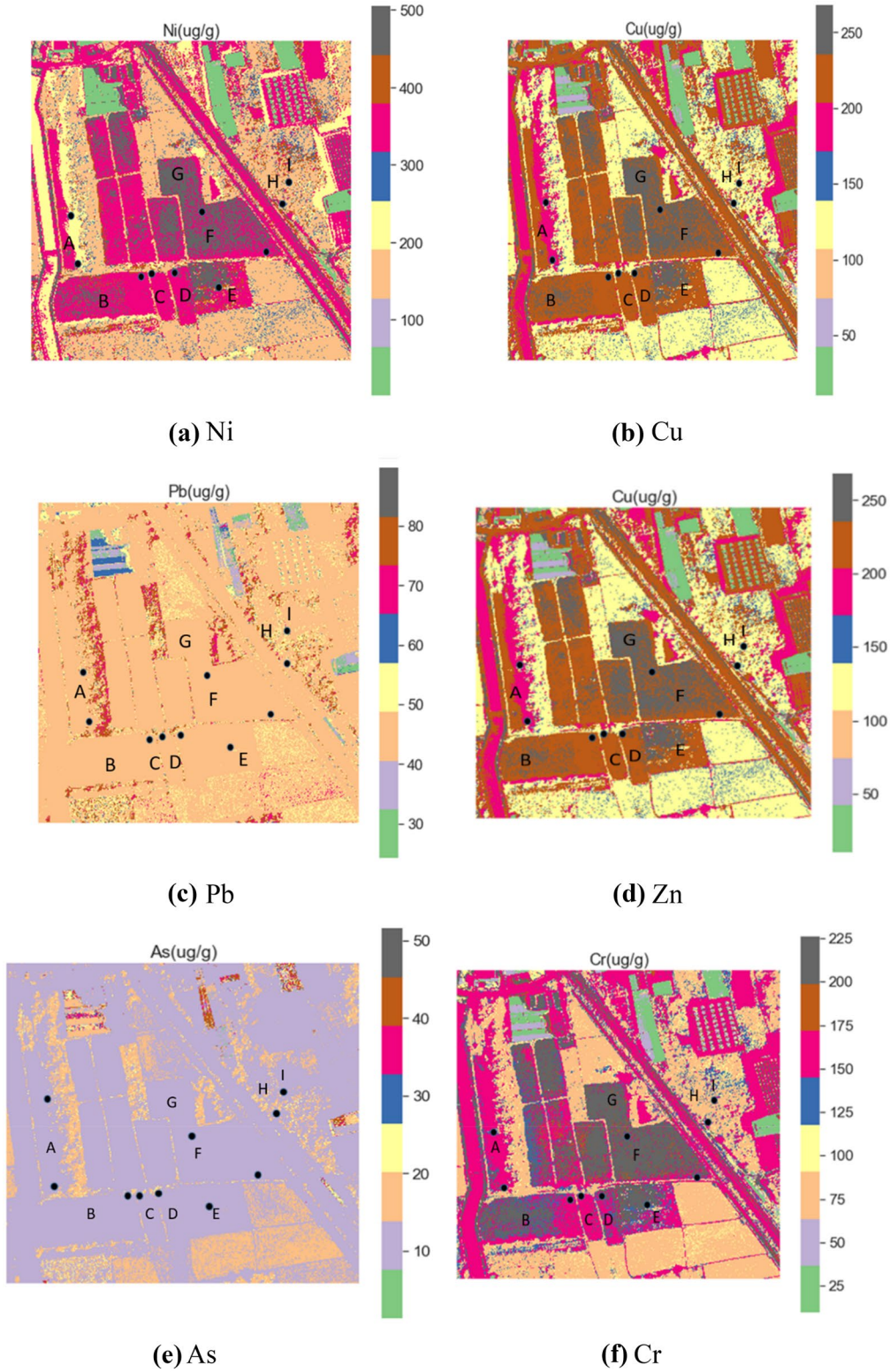


Fig. 8 Spatial patterns of the heavy metals

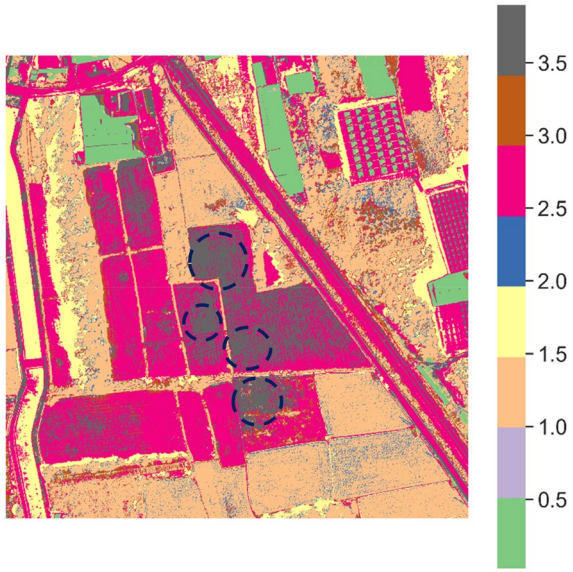


Fig. 9 Comprehensive hazard analysis of heavy metals

$$\sum_i \frac{x_i}{x_{i,s}} \tag{11}$$

where

x_i = the predicted concentration of the ith metal in each pixel

$x_{i,s}$ = the monitoring standard of the ith metal (see Table 1)

Conclusion

We use literature and on-site data for modeling. The model shows that the method proposed in this research can effectively distinguish between polluted and non-polluted areas, but the lack of local samples limits the development and application potential of the model. Therefore, this study considers the establishment of a local soil database in the future to increase the accuracy of the model. This study reveals that in the case of a small sample, random forest has better prediction results. The 550 band in the blue light spectrum has a decisive influence on the prediction results of the model, and the long band of 735 is likely to cause variation in the prediction results of the model. Random forest can effectively predict the high concentrations of Cu, Ni, and Cr in the study plots. The low concentrations of Pb and As are not very good because of the small variation of the sample. The prediction effect of Random Forest or BPNN is not good. When screening soil characteristic variables in this study, it can be compared with the feature importance of the model at the same time, which can improve the prediction accuracy of the model, and design a random forest algorithm combined with other algorithms to optimize parameters.

In this study, it is difficult to obtain more spectral data of the band due to the limitation of the instrument. If more spectral band data can be captured in the future, the accuracy of proposed model will be improved. In addition, the limitation of the study is using digital images and filters that not able to display dynamic movement of vegetation growing and elements moving in soil. The model's accuracy would be improved in the future, if the study would consider the composition of the soil itself to predict heavy metal contents. Establishment of local soil database is also an initiative for increasing the accuracy of the model.

The proposed prediction model can help decision makers to distinguish effectively between contaminated lands and non-contaminated lands. The spatial patterns of Cu, Ni, and Cr can be accurately figured out by proposed model. Using this model, high pollution areas can be quickly found, and following investigation and remediation work can be carried out.

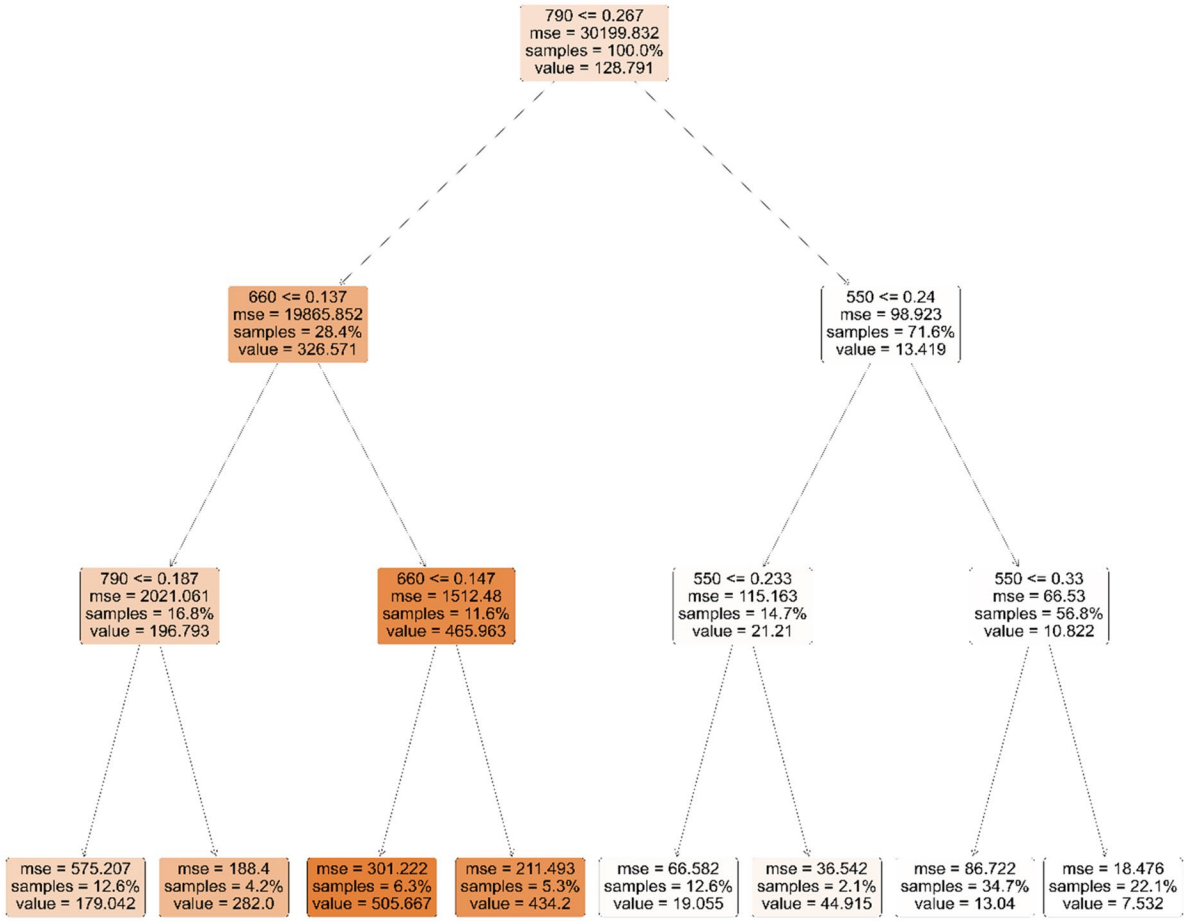
Funding The authors received financial support from the Environmental Protection Administration in Taiwan under Contract No. 109C003942.

Data availability All data generated or analyzed during this study are included in this published article [and its supplementary information files]. The raw data is available on request from the authors.

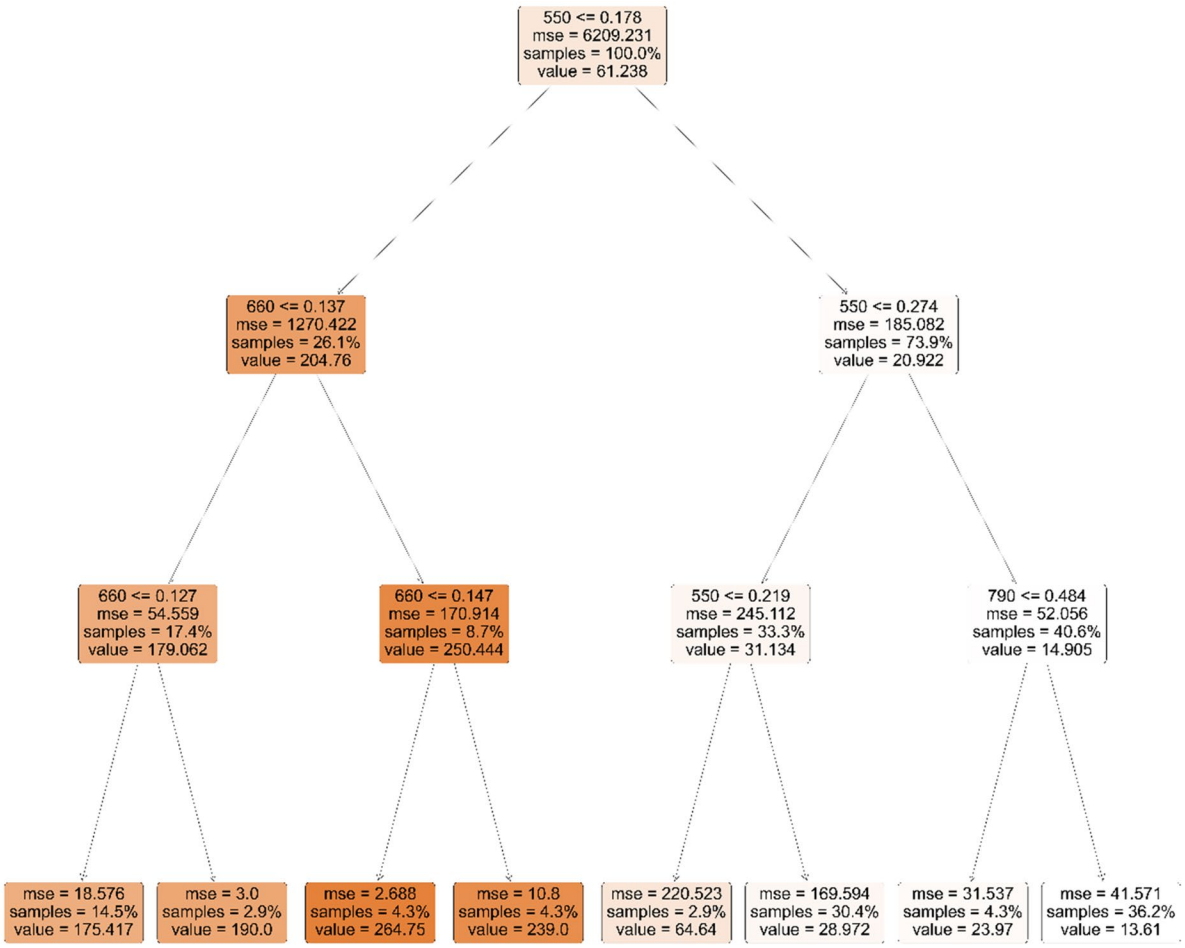
Declarations

Conflict of interest The authors declare no competing interests.

Appendix 1. Random forest-based prediction for Ni



Appendix 2. Random forest-based prediction for Cu



References

Albanese, S., De Vivo, B., Lima, A., & Cicchella, D. (2007). Geochemical background and baseline values of toxic elements in stream sediments of Campania region (Italy). *Journal of Geochemical Exploration, 93*(1), 21–34. https://doi.org/10.1016/j.gexplo.2006.07.006

Antoniadis, V., Shaheen, S., Levizou, E., Shahid, M., Niazi, N., Vithanage, M., et al. (2019). A critical prospective analysis of the potential toxicity of trace element regulation limits in soils worldwide: Are they protective concerning health risk assessment? - A review. *Environment International, 127*, 819–847. https://doi.org/10.1016/j.envint.2019.03.039

Bell, R., & Russell, C. (2002). Environmental policy for developing countries. *Issues in Science and Technology, 18*(3), 63–70. Retrieved May 28, 2021, from http://www.jstor.org/stable/43314167. Accessed 15 January 2021.

Briffa, J., Sinagra, E., & Blundell, R. (2020). Heavy metal pollution in the environment and their toxicological effects on humans. *Heliyon, 6*(9), e04691. https://doi.org/10.1016/j.heliyon.2020.e04691 Buaisha, M., Balku, S., & Özalp-Yaman, S. (2020). Heavy metal removal investigation in conventional activated sludge systems. *Civil Engineering Journal, 6*(3), 470–477.

Bundschuh, J., Litter, M., Parvez, F., Román-Ross, G., Nicolli, H., Jean, J., et al. (2012). One century of arsenic exposure in Latin America: A review of history and occurrence from 14 countries. *Science of the Total Environment, 429*, 2–35. https://doi.org/10.1016/j.scitotenv.2011.06.024

Chowdhury, S., Mazumder, M., Al-Attas, O., & Husain, T. (2016). Heavy metals in drinking water: Occurrences, implications, and future needs in developing countries. *Science of the Total Environment, 569–570*, 476–488. https://doi.org/10.1016/j.scitotenv.2016.06.166

Díez, M., Simón, M., Martín, F., Dorronsoro, C., García, I., & Van Gestel, C. (2009). Ambient trace element background concentrations in soils and their use in risk assessment. *Science of the Total Environment, 407*(16), 4622–4632. https://doi.org/10.1016/j.scitotenv.2009.05.012

Hafez, Y., & Awad, E. (2016). Finite element modeling of radon distribution in natural soils of different geophysical regions. *Cogent Physics*, *3*(1). https://doi.org/10.1080/23311940.2016.1254859

Hu, P., Huang, J., Ouyang, Y., Wu, L., Song, J., Wang, S., et al. (2013). Water management affects arsenic and cadmium accumulation in different rice cultivars. *Environmental Geochemistry and Health, 35*(6), 767–778. https://doi.org/10.1007/s10653-013-9533-z

Jia, X., Cao, Y., O'Connor, D., Zhu, J., Tsang, D. C., Zou, B., & Hou, D. (2021). Mapping soil pollution by using drone image recognition and machine learning at an arsenic-contaminated agricultural field. *Environmental Pollution, 270*, 116281.

Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. Proceedings Of ICNN'95 - International Conference On Neural Networks. https://doi.org/10.1109/icnn.1995.488968

Lai, H., Hseu, Z., Chen, T., Chen, B., Guo, H., & Chen, Z. (2010). Health risk-based assessment and management of heavy metals-contaminated soil sites in Taiwan. *International Journal of Environmental Research and Public Health, 7*(10), 3595–3614. https://doi.org/10.3390/ijerph7103596

Lal, R. (2000). Soil management in the developing countries. *Soil Science, 165*(1), 57–72. https://doi.org/10.1097/00010694-200001000-00008

Lan, Y., Huang, Z., Deng, X., Zhu, Z., Huang, H., Zheng, Z., ... & Tong, Z. (2020). Comparison of machine learning methods for citrus greening detection on UAV multispectral images. *Computers and Electronics in Agriculture, 171*, 105234.

Lin, Y., Teng, T., & Chang, T. (2002). Multivariate analysis of soil heavy metal pollution and landscape pattern in Changhua county in Taiwan. *Landscape and Urban Planning, 62*(1), 19–35. https://doi.org/10.1016/s0169-2046(02)00094-4

Myint, H. (1973). *The economic of the developing countries*. Hutchinson University library.

Qin, G., Niu, Z., Yu, J., Li, Z., Ma, J., & Xiang, P. (2021). Soil heavy metal pollution and food safety in China: Effects, sources and removing technology. *Chemosphere, 267*, 129205. https://doi.org/10.1016/j.chemosphere.2020.129205

Qun'ou, J., Lidan, X., Siyang, S., Meilin, W., & Huijie, X. (2021). Retrieval model for total nitrogen concentration based on UAV hyper spectral remote sensing data and machine learning algorithms–A case study in the Miyun Reservoir, China. *Ecological Indicators, 124,* 107356.

Rajaganapa, V., Xavier, F., Sreekumar, D., & Mandal, P. (2011). Heavy metal contamination in soil, water and fodder and their presence in livestock and products : A review. *Journal of Environmental Science and Technology, 4*(3), 234–249. https://doi.org/10.3923/jest.2011.234.249

Rodríguez Eugenio, N., McLaughlin, M., & Pennock, D. (2018). *Soil pollution*. Food and Agriculture Organization of the United Nations.

Sutherland, R. A. (2000). Bed sediment-associated trace metals in an urban stream, Oahu, Hawaii. *Environmental Geology, 39*(6), 611–627.

Zwolak, A., Sarzyńska, M., Szpyrka, E., & Stawarczyk, K. (2019). Sources of soil pollution by heavy metals and their accumulation in vegetables: A review. *Water, Air, & Soil Pollution*, *230*(7). https://doi.org/10.1007/s11270-019-4221-y