Environmental Pollution 250 (2019) 601–609

Contents lists available at ScienceDirect

Environmental Pollution

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

A methodological framework for identifying potential sources of soil heavy metal pollution based on machine learning: A case study in the Yangtze Delta, China☆

Xiaolin Jia [a], Bifeng Hu [b, c], Ben P. Marchant [d], Lianqing Zhou [a], Zhou Shi [a, *], Youwei Zhu [e]

[a] Institute of Agricultural Remote Sensing & Information Technology Application, College of Environmental and Resource Sciences, Zhejiang University, Hangzhou, Zhejiang, 310058, China
[b] Unité de Recherche en Science du Sol, INRA, Orléans, 45075, France
[c] InfoSol, INRA, US 1106, Orléans, 45075, France
[d] British Geological Survey, Keyworth, Nottinghamshire, NG12 5GG, UK
[e] Zhejiang Management Bureau of Planting, Hangzhou, Zhejiang, 310020, China

ARTICLE INFO

Article history:
Received 11 January 2019
Received in revised form
28 March 2019
Accepted 9 April 2019
Available online 12 April 2019

Keywords:
Heavy metal pollution
Source identification
Potentially polluting enterprises
Multinomial naive bayesian methods
Bivariate local Moran's I analysis

ABSTRACT

It is a great challenge to identify the many and varied sources of soil heavy metal pollution. Often little information is available regarding the anthropogenic factors and enterprises that could potentially pollute soils. In this study we use freely available geographical data from a search engine in conjunction with machine learning methodologies to identify and classify potentially polluting enterprises in the Yangtze Delta, China. The data were classified into 31 separate and four integrated industry types by five different machine learning approaches. Multinomial naive Bayesian (NB) methods achieved an accuracy of 87% and Kappa coefficient of 0.82 and were used to classify the geographic data from more than 260,000 enterprises. The relationship between the different industry classes and measurements of soil cadmium (Cd) and mercury (Hg) concentrations was explored using bivariate local Moran's I analysis. The analysis revealed areas where different industry classes had led to soil pollution. In the case of Cd, elevated concentrations also occurred in some areas because of excessive fertilization and coal mining. This study provides a new approach to investigate the interaction between anthropogenic pollution and natural sources of soil heavy metals to inform pollution control and planning decisions regarding the location of industrial sites.

1. Introduction

Rapid economic and industrial development has led to the accumulation of soil heavy metal elements of impacted sites across the world (Facchinelli et al., 2001; Marchant et al., 2011; Hu et al., 2019). Heavy metals generally have persistent bioavailability, long residence times (commonly exceeding decades), and often low concentration thresholds indicate toxicity (Jiang et al., 2017). The excessive accumulation of heavy metals can hence disrupt the usual biochemical processes which occur in soils, leading to deterioration of soil quality, reduced agricultural productivity and quality and human health risks (Dudzik et al., 2010; Zawadzka and Łukowski, 2010; Marchant et al., 2017). The 16.1% of soil samples are contaminated with heavy metals and therefore detailed studies of soil contamination in China are required (Hu et al., 2017b).

Human enterprises such as industry, transportation and agriculture can be the source of substantial quantities of soil heavy metal elements (Facchinelli et al., 2001). According to the Statistical Yearbook of China in 2014, the number of registered enterprises in China was approximately 22.6 million, and the bankrupt and newly established enterprises were, in combination, approximately 30% of the all enterprises in China. It is extremely difficult to make timely investigations and reports of the pollution effects of different enterprises across large regions using traditional methods, especially when the region is large and dispersed. The traditional source

☆ This paper has been recommended for acceptance by Dr. Yong Sik Ok.

* Corresponding author. Institute of Agricultural Remote Sensing and Information Technology Application, College of Environmental and Resource Sciences, Zhejiang University, Yuhangtang Road 866, Hangzhou, 310058, Zhejiang, China.
E-mail addresses: 11814013@zju.edu.cn (X. Jia), bifeng.hu@inra.fr (B. Hu), benmarch@bgs.ac.uk (B.P. Marchant), lianqing@zju.edu.cn (L. Zhou), shizhou@zju.edu.cn (Z. Shi), 13018941333@163.com (Y. Zhu).

https://doi.org/10.1016/j.envpol.2019.04.047
0269-7491/

apportionment methods mainly include principal component analysis (PCA), isotope ratio analysis, positive matrix factorization (PMF) and stochastic models (Qu et al., 2013; Luo et al., 2014; Wang et al., 2016). For example, Hu and Cheng (2013) analyzed seven environment variables relevant to the soil heavy metal pollution using stochastic models, and Ma et al. (2018) researched the major potential source of soil heavy metals and human health risk using PCA in high population density area. These methods analyze the contribution of different sources to soil heavy metal pollution, but ignore the spatial distribution and characteristics of these sources (Duodu et al., 2017; Guan et al., 2018). Furthermore, the model mechanisms and data collection requirements of diffusion models of source apportionment for soil heavy metals are very complex, which is not convenient for wide uptake across large-scale regions. Exhaustive information regarding the location and type of enterprises within a region is rarely available.

In this study, we use freely available geographic information from a search engine to build an inventory of enterprises within the Yangtze Delta region of China. This geographic data does not specify the type of industry. We therefore survey a subset of the enterprise locations and form a training dataset of enterprise types. We test five different machine learning approaches to build a classifier of enterprise type and apply the best performing method to the full geographic dataset. We illustrate how the derived dataset might be utilized by using the bivariate local Moran's I method to analyze the spatial correlation between these enterprises and elevated soil metal concentrations and thus provide effective guidance and assistance for the management and control of these anthropogenic sources of pollution (Piroonsup and Sinthupinyo, 2017; Salles et al., 2017).

2. Materials and methods

2.1. Study area

The study area (27° 02′-31° 11′ N, 118° 01′-123° 10′ E) is located in the Yangtze Delta of China, which covers 105,500 km^2 and has a population of 55.9 million (see Fig. 1). The study area possesses a typical subtropical monsoon climate, which is mild and humid with annual average temperature of 16.5 °C and annual average precipitation of 1575 mm. The western, eastern and southern parts of study area are mainly red soil and yellow soil, and the southeast coastal and northern parts are mainly paddy soil. The industries in study area mainly include textile industry, chemical industry, and metalwork industry. The study area is one of the most developed regions in China and the concentrations of soil heavy metals are also remarkably high. According to Soil Pollution Condition Investigation Communique in 2013, the proportion of samples contaminated by the chromium (Cr), lead (Pb), cadmium (Cd), mercury (Hg), arsenic (As) elements in study area were 0.9%, 0.2%, 15.6%, 10.9% and 1.0%, respectively. This study mainly focused on the source apportionment of Cd and Hg.

2.2. Soil sampling

A total of 14,801 topsoil samples were collected from the study area in 2013 by the method of systematic grid sampling (1 km × 1 km). Each soil sample was the bulked combination of five subsamples from five locations within 5 m at a depth of 0–20 cm. Fresh soil samples were transported to the laboratory, air-dried, ground, and passed through a 2 mm sieve. Soil pH was measured in H_2O with the soil and solution ratio of 1:2.5 (m/v) using the glass electrode method. The Cd element in soils was digested by HF–HNO_3–$HClO_4$ and measured by an inductively coupled plasma-mass spectrometer (ICP-MS, Agilent 7500a, Palo Alto, CA, USA). The Hg element in soils was digested by HNO_3–HCl and determined by an atomic fluorescence spectrometer (Atomic Fluorescence Spectrometry, AFS). For the dependability of results, blank control, duplicate samples, and standard reference soils were used in chemical analysis.

The degree of pollution in each soil sample was calculated using the single pollution index (SPI), which was calculated based on the national standard values of different soil heavy metal elements as the evaluation criterion (Hu et al., 2017a). The equation is defined as: $SPI_i = \frac{C_i}{S_i}$, where C_i is the measured concentration of heavy metal i, and S_i is the national standard values. The SPI_i is classified as safety ($SPI_i \leq 1.0$), slight contamination ($1.0 < SPI_i \leq 2.0$), mild contamination ($2.0 < SPI_i \leq 3.0$), moderate contamination ($3.0 < SPI_i \leq 5.0$), and severe contamination ($SPI_i > 5.0$).

2.3. Data collection

Information including the latitude and longitude, potential contaminants, enterprise name and industrial category of 7643 potentially polluting enterprises was collected by field investigation. The dataset included 31 industrial categories. Almost 80% of the sites belonged to textile industry (30%), chemical industry (29%) and metalwork industry (19%). The other 28 industrial categories accounted for only 22% of the whole dataset, and the proportion of any single industry type was never more than 4%. The data classified according to the complete set of 31 industrial types is referred to as the separated dataset. We also formed an integrated dataset where the textile industry, chemical industry and metalwork industry classes were retained whilst the remainders were combined into a single class.

Google search API data consisting of latitude and longitude and enterprise name was acquired for 264,098 sites using the keyword 'enterprise'. This search information did not include industry type which is likely to be a critical factor controlling the degree of soil pollution. Machine learning methodologies were therefore adopted to classify the industry types (as recorded in the field survey) using the Google search data.

2.4. Industrial classification

The main steps in performing classification of industry types, based primary on the enterprises name, were: 1) Word segmentation. The word segmentation, based on a hidden Markov model divided the text into words and the word corpus originating from the training (i.e. field investigation) samples was used for the segmentation of the unlabelled samples. 2) Feature vectorization. The feature vectorization consisted of the feature extraction and the feature selection. Feature extraction is required to remove noise, stop words and other irrelevant text and then present the text in vector form to the classification models. Feature selection leads to improved classification efficiency and reduces the computational complexity. The information gain method based on entropy was used to process this step in this study. The results were analyzed and evaluated by using the Kappa coefficient, which is used to measure observer agreement for categorical data, and large Kappa coefficients indicate an accurate model (Landis and Koch, 1977). 3) Classification modelling. The classification models considered were Support Vector Machine (SVM), naive Bayesian (NB) and Artificial Neural Network (ANN) algorithm.

The SVM algorithm seeks the best compromise between the model complexity and the learning ability based on the principle of structural risk minimization (Keerthi and Lin, 2003; Ma et al., 2017). The common kernel functions are linear, polynomial, sigmoid and radial basis function (RBF). The RBF kernel nonlinearly maps samples into a higher dimensional space and handles the case when the relation between class labels and attributes is nonlinear. The linear

kernel is a special case of RBF, which is a better choice in the case of the large number of features. The polynomial kernel has a large number of hyperparameters, which greatly increases the complexity of model selection. The sigmoid kernel behaves like RBF and is not valid under some parameters. In general, the RBF and linear kernels were reasonable first choices in this study. NB is a classification method based on Bayes' theorem and independent assumption of feature conditions (Khalil, 2018). First, for a given training dataset, the joint probability distribution of the feature and classification (prior probability model) is learned based on independent assumption of feature conditions. Then, for a given input x (feature), the output y (classification) with the greatest posterior probability is obtained using the maximum likelihood estimation function and Bayes' theorem. Multinomial NB and Bernoulli NB are the most commonly used models and respectively belong to bag-of-words model and set-of-words model. The ANN algorithm simulates structural and functional characteristics of biological neural network and is used for pattern analysis, signal processing and so on (Sun et al., 2018). It has the advantages of self-learning, nonlinear mapping, and flexible network structure. Details of the industrial classification procedure are illustrated in Fig. 2.

2.5. Enterprise density

Kernel Density (KD) was used to create a smoothed surface of industry distribution by Google search API data (Deng et al., 2019). Results were interpreted as density of enterprises per square kilometer. The KD is given by:

$$\rho(s) = \frac{1}{nr}\sum_{i=1}^{n} k\left(\frac{d_{is}}{r}\right) \quad (1)$$

where $\rho(s)$ is the density at location s, r is the search radius (bandwidth), n is the number of sampling points in the search radius range, and k is the weight of distance d_{is} between a point i and location s and usually defined by the quartic kernel function.

2.6. Bivariate spatial correlation analysis

Traditional statistical analysis methods usually focus on statistical relationship between different variables recorded at the same site. However, pollution from an enterprise can potentially extend over a wider area. To overcome this gap, bivariate spatial correlation analysis was conducted to identify spatial association patterns of the industry type and soil pollution data. The study area was divided into nearly 5000 (5 km × 5 km) grid cells, and the bivariate local Moran's I (I^{ab}) was applied for the spatial analysis of the grid data (Equation (2)) (Wu et al., 2019).

$$I^{ab} = X_i^a \sum_{j=1,\, j\neq i}^{n} w_{ij} X_j^b \quad (2)$$

where X_i^a and X_j^b respectively is the value of the variable a and b at location i and j; and w_{ij} is defined as the spatial weight matrix based on a distance weighting between locations i and j. When I^{ab} is significantly positive or negative, it shows that the variable a at the grid i is observably correlated with the variable b in the adjacent area; if not, it means that there be no obvious correlation between them.

3. Results and discussion

3.1. Heavy metal and enterprise contaminant statistics

The descriptive statistics regarding the concentration of Cd and Hg in soil are shown in Table 1. The concentration ranges of Cd and Hg were respectively 0.0036–114 and 0.008–7 mg kg^{-1}. The mean concentrations of Cd and Hg were respectively 0.3 and 0.2 mg kg^{-1}, which are both higher than the soil background concentrations in the study area (0.07 and 0.09 mg kg^{-1}) and in China (0.10 and 0.07 mg kg^{-1}) (Hu et al., 2017a). The coefficient of variation (CV) of Cd and Hg with the values of 367% and 100% indicated the presence of extremely large concentrations of each element possibly due to anthropogenic activities.

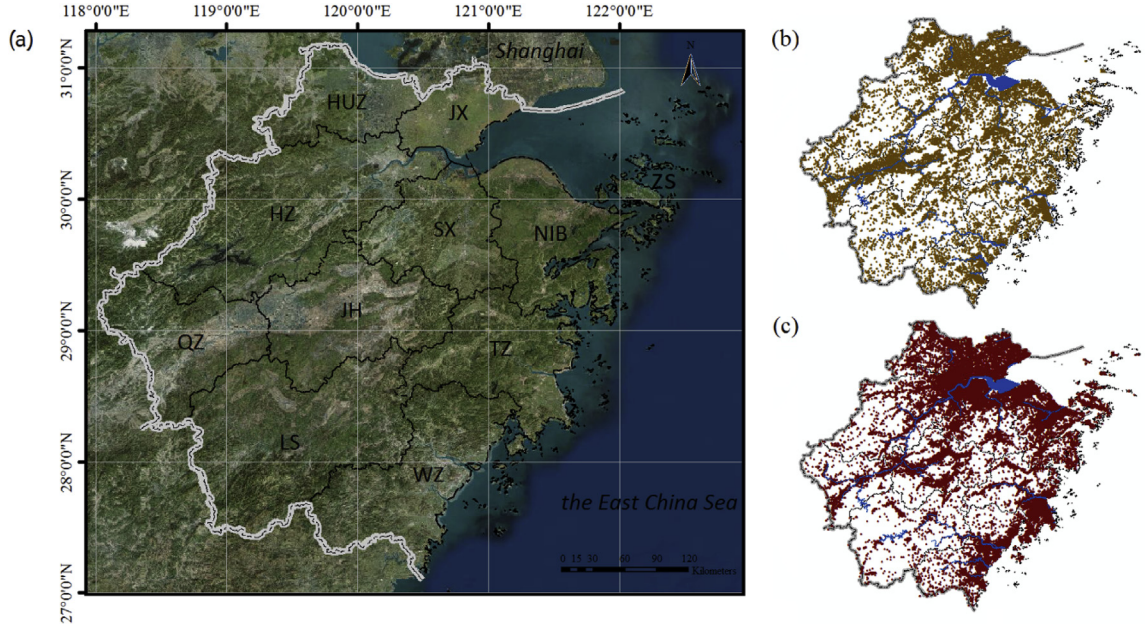

Fig. 1. Maps showing the location of the study area, soil sampling and enterprise sites in the Yangtze Delta of China. a: HUZ, HZ, QZ, LS, JH, SX, JX, ZS, NIB, TZ, WZ were respectively the English abbreviations of the 11 provincial cities in the study area, b: the yellow points and the blue polygon respectively represented the 14,801 soil samples and the river system, c: the red points represented the 264,098 enterprises. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

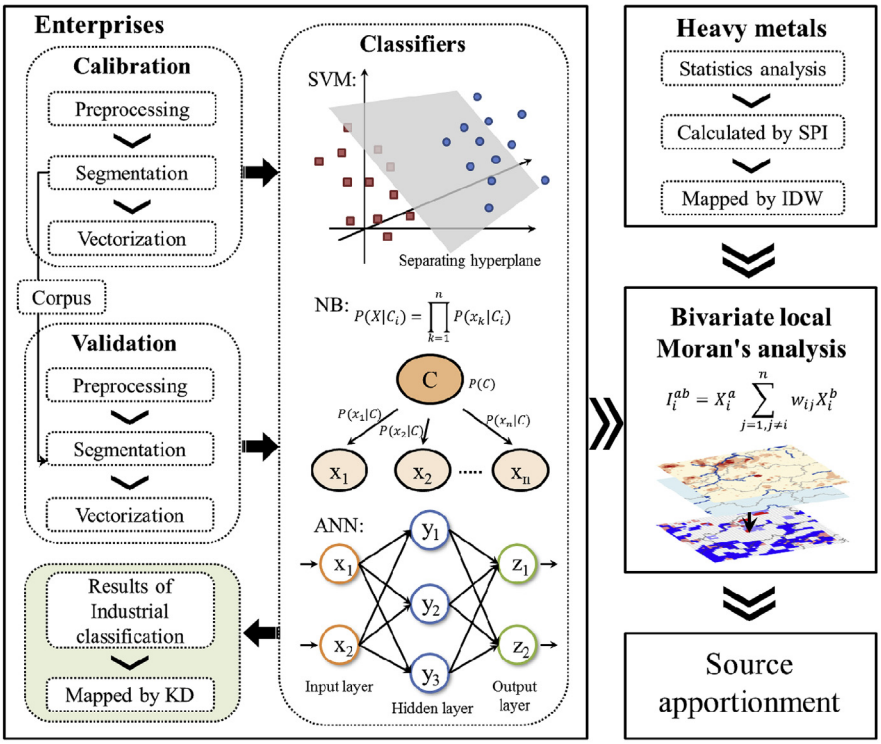

Fig. 2. Workflow of the source apportionment in this study. KD: the Kernel Density method, SPI: the single pollution index, IDW: the Inverse Distance Weighted method, SVM: the Support Vector Machine method, NB: the naive Bayesian method, ANN: the Artificial Neural Network method.

Table 1
Descriptive statistics for the Cd and Hg concentrations in soils.

Element	Mean	Median	SD	Skew	Min	Max	CV	SBC_1	SBC_2
Cd (mg kg^{-1})	0.3	0.18	1.1	88.2	0.0036	114	367%	0.07	0.10
Hg (mg kg^{-1})	0.2	0.11	0.2	8.2	0.008	7	100%	0.09	0.07

SD: the standard deviation; CV: the coefficient of variation; SBC_1: the soil back concentrations in the study area; SBC_2: the soil back concentrations in China.

According to results of the field survey of 7643 enterprise contaminants (Table 2), the proportions of Cd contaminant in the textile industry, metalwork industry and chemical industry were respectively 1%, 18% and 10%, meanwhile the proportions of Hg contaminant were respectively 1%, 12% and 17%. In the three industries, Cd and Hg contaminants were respectively produced mainly by the metalwork industry and the chemical industry. Based on the statistics of the contaminants, the main contaminant of the textile industry, metalwork industry and chemical industry was Cr. Cr element was basically pollution-free in soils, therefore Cr element wasn't considered in this study.

3.2. Industrial classification of enterprises

For analyzing the classification accuracy of different machine learning models, the training samples were divided into a calibration dataset (1148 samples) and a validation dataset (6495 samples). The radial basis function kernel and linear kernel were used within the SVM classification models. Multinomial NB and Bernoulli NB classified enterprises by adopting different strategies for calculating the likelihood probability of characteristics. The ANN model was a simple network model with only one hidden layer. The prediction results using different classification models are shown in Table 3. These five models had good predictive ability with high accuracy on both the separated and integrated datasets. The average accuracies of prediction results in calibration and validation dataset were 97% and 84% respectively. Overall, the accuracy of models using integrated samples was superior to those using separated samples. SVM, NB and ANN were improved by 2–3%, 3% and 4% in validation dataset, respectively. The SVM with linear kernel performed best on the calibration dataset with accuracies of 99% and 99% for the calibration dataset. However, by the comprehensive consideration of the results in different datasets, Multinomial NB was chosen to classify the enterprises since it had the highest accuracies of 87% in the integrated validation dataset.

Table 2
Proportion of each heavy metal contaminants produced by enterprises.

	Cr	Pb	Cd	Hg	As
Textile industry	94%	2%	**1%**	**1%**	2%
Metalwork industry	55%	15%	**18%**	**12%**	0%
Chemical industry	40%	20%	**10%**	**17%**	13%

For Multinomial NB model, the numbers of enterprise samples predicted correctly in validation dataset, of which industrial classifications were textile industry, chemical industry, metalwork industry and the other industry, were respectively 178, 274, 348 and 193 (Table 4). The average values of the prediction classification accuracy and the method classification accuracies in validation dataset were respectively 86% and 88%. The method classification accuracies in validation dataset were, from high to low, textile industry, the other industry, metalwork industry and chemical industry, respectively. The prediction classification accuracies in validation dataset followed the order: metalwork industry>chemical industry>textile industry>the other industry. The Kappa coefficient was 0.82 that meant that the predicted results of industrial classification of polluting enterprises by Multinomial NB

Table 3
Correct rates of different industry classification models in calibration and validation datasets.

Dataset	Separation					Integration				
	SVM_a	SVM_b	NB_a	NB_b	ANN	SVM_a	SVM_b	NB_a	NB_b	ANN
Calibration	98%	99%	95%	92%	99%	98%	99%	94%	94%	99%
Validation	82%	84%	84%	82%	81%	85%	86%	**87%**	85%	85%

SVM_a and SVM_b: the Support Vector Machine model respectively with radial basis function kernel and linear kernel; NB_a and NB_b: the naive Bayesian model respectively using the Multinomial and Bernoulli theorem; ANN: the Artificial Neural Network model; Separation: the 31 industrial classifications; Integration: the 4 industrial classifications.

Table 4
Comparison for industrial classification results of Multinomial NB model with the observed results.

Predicted	Actual					
	Textile industry	Chemical industry	Metalwork industry	The other industry	Total	Method Classification Accuracy
Textile industry	**178**	3	1	3	185	96%
Chemical industry	4	**274**	11	41	330	83%
Metalwork industry	17	22	**348**	26	413	84%
The other industry	7	16	4	**193**	220	88%
Total	206	315	364	263	1148	/
Prediction Classification Accuracy	86%	87%	96%	73%	/	**Kappa Coefficient: 0.82**

were almost identical with the actual results.

3.3. Applicability analysis of classification models

By comparing the classification results of separated and integrated samples, the model accuracies of SVM, NB and ANN were respectively improved by 0%, -1-2% and 0% in calibration dataset and 2–3%, 3% and 4% in validation dataset after data integration. SVM with linear kernel and Multinomial NB, which were the models with the highest accuracies in validation dataset, had an improvement of 2% and 3% after data integration. The SVM approach was least sensitive to the number of industry classes and hence more widely applicable. The classification accuracies of SVM using the linear kernel were comparable to that using the RBF kernel. Apparently, when the number of features was large, the nonlinear mapping couldn't improve the performance of classification models. The linear kernel was more appropriate for industrial classification than the RBF kernel in this study. The NB approach required a prior probability value in the classification process and hence was more sensitive to the distribution of categories of data than the other models. In this study, the training samples and unlabelled samples had similar distributions of industrial types since they were collected from the same research area. The spatial correction between the soil heavy metals and the main industries were analyzed, using only the textile industry, metalwork industry and chemical industry classes. Therefore, Multinomial NB was chosen for the industrial classification of unlabelled samples, as it had the highest accuracy of 87%. However, SVM with linear kernel had the best applicability ability. In the future work, when try to apply the classification model on the national scale, it is necessary to consider the applicability ability of models and adopt SVM with linear kernel.

3.4. Spatial distribution of heavy metals and enterprises

The search engine data classified to either the textile industry, metalwork industry or chemical industry were retained, accounting for 10%, 28% and 42% of the total content, respectively. The spatial distribution of enterprise density and soil heavy metal pollution degree are shown in Fig. 3. The textile industry, metalwork industry and chemical industry was distributed mainly in the eastern part of study area and near rivers and lakes. The number of the enterprises belonged to textile industry was less than that of the other industries and mainly distributed in the JX district. The enterprises in metalwork industry and chemical industry had the similar distribution, which were mainly in the HZ, NIB, WZ, TZ districts. The region seriously contaminated by Cd was located in the QZ and HZ districts, and Hg contaminated region was mainly located in the SX and NIB districts. The WZ district included Cd and Hg pollution in soils, where the contamination degree and area was relatively low.

3.5. Source apportionment of soil heavy metal pollution

The spatial correlation degree between the different element concentrations and the different enterprises as calculated from the bivariate local Moran's I analysis is shown in Fig. 4(a–f). According to relevant researches on the response of industrial spatial pattern to ecological environment, the following conclusions were presented (Rigina, 2002; Verhoef and Nijkamp, 2002; Grazi et al., 2007): 1) High-high and high-low indicate areas with high heavy metal concentrations. In the former case this is likely to be the result of pollution from the enterprises. This area has a high level of industrial development and a certain scale effect based on industrial agglomeration, whereas the rapid and large-scale industrial development inevitably leads to the aggravation of environmental pollution and greatly restricts the sustainable development of this area. In the latter case it is more likely to result from the other environmental factors such as agricultural production or soil parent material. 2) Low-high and low-low indicate uncontaminated areas. Low-high area is mainly distributed in the periphery of high-high area. Due to spatial neighborhood and spillover effect, the industry developed rapidly and no soil pollution resulted in this area. With the improvement of industrial agglomeration, it is likely to be a high-speed growth area of pollution in the future.

According to the results of bivariate spatial correlation analysis, Cd pollution in soils was mainly unrelated to the enterprises and located in the QZ and HZ districts, while soil Hg pollution was seriously affected by enterprises that was mainly distributed in the JX, SX, NIB and WZ districts. Considering the Cd pollution, the QZ and HZ districts mainly had high-low area and a few high-low areas were sparsely located in the TZ, LS and WZ districts, meanwhile the JH, SX, WZ and TZ districts contained a small number of scattered high-high areas. In the case of Cd pollution, the textile industry led

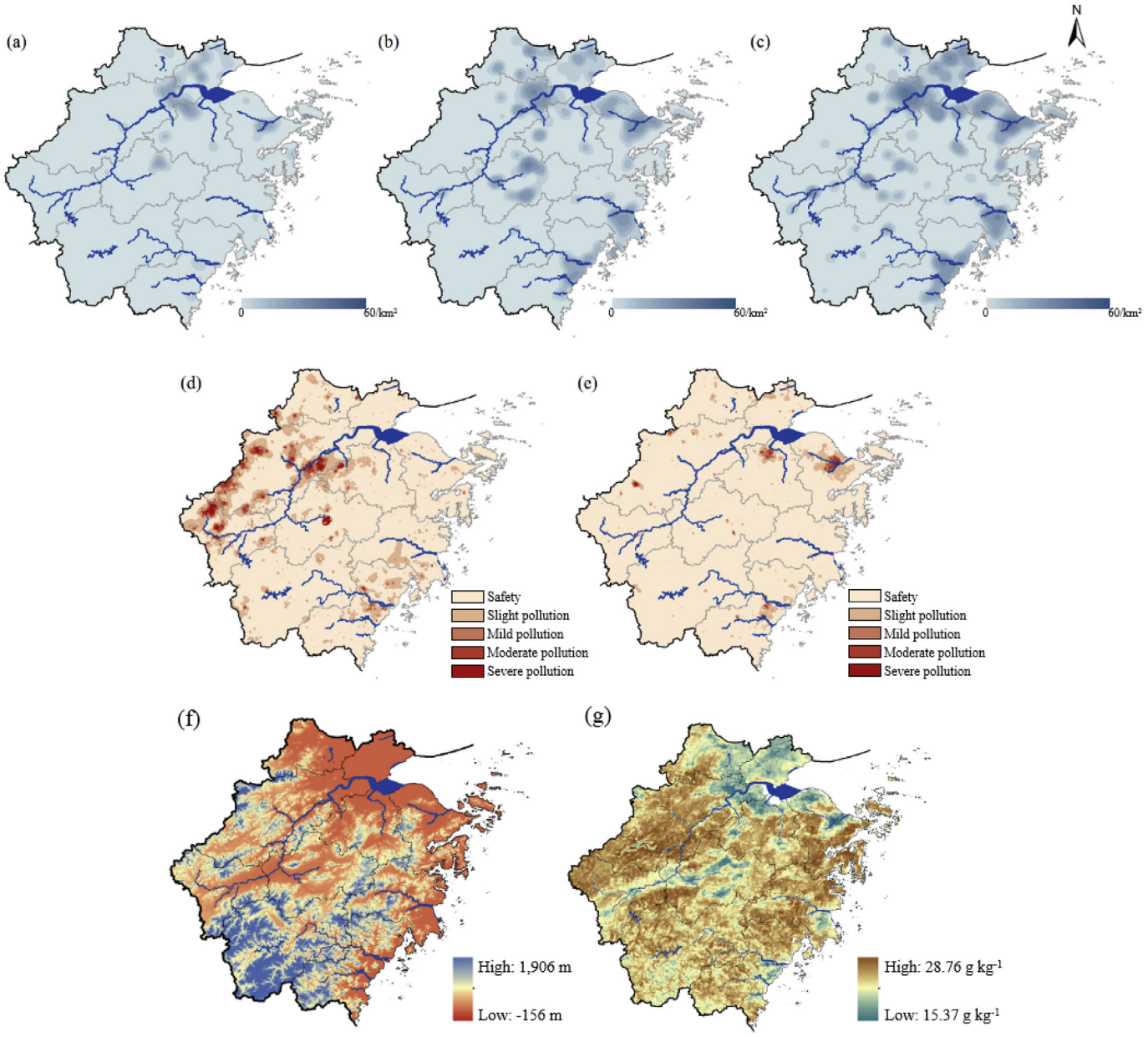

Fig. 3. Spatial distribution of enterprise density and soil properties. a: textile industry, b: metalwork industry, c: chemical industry, d: Cd element pollution degree, e: Hg element pollution degree, f: digital elevation model (DEM), e: soil organic matter (SOM). The density of pollution enterprises and the soil properties were respectively mapped by the Kernel Density (KD) method and the Inverse Distance Weighted (IDW) method.

to almost no pollution in the TZ district and the chemical industry had almost no pollution in the JH district. The metalwork industry caused Cd pollution more seriously than the other industries.

In the SX district, the Hg pollution was caused mainly by the textile industry and chemical industry, and in the WZ district by the metalwork industry and chemical industry. Moreover, the chemical industry had the largest high-high area in Hg pollution. The average high-low area of Cd in the different enterprise analysis was 4277 km^2, which was 4% of the whole study area, while the average area of Hg was 107 km^2, 0.1% of the study area. The areas of Cd pollution mainly caused by the textile industry, metalwork industry and chemical industry were respectively 908, 1575 and 1162 km^2, while the areas of Hg pollution were respectively 1,442, 1717 and 1904 km^2. The high-high distribution of Hg was relatively agglomerated compared with Cd.

The four scenarios of spatial correlation between heavy metals and enterprises were shown in the coordinate system (Fig. 4 g-h). Because of the similar scatter plots of three industries, the textile industry was cited to explain this result mainly. According to Fig. 4(g and h), the distribution of the different scenarios had significantly differences. The boundary that denoted relatively high and low heavy metal pollution potential condition was 1, which was consistent with the index value defining heavy metal contamination calculated by SPI. The boundary that divided relatively high and low enterprise density were between 0 and 25.

3.6. Non-enterprise soil pollution sources

In the late 1970s and early 1980s, the background values of heavy metal elements in soils of the study area were studied by the

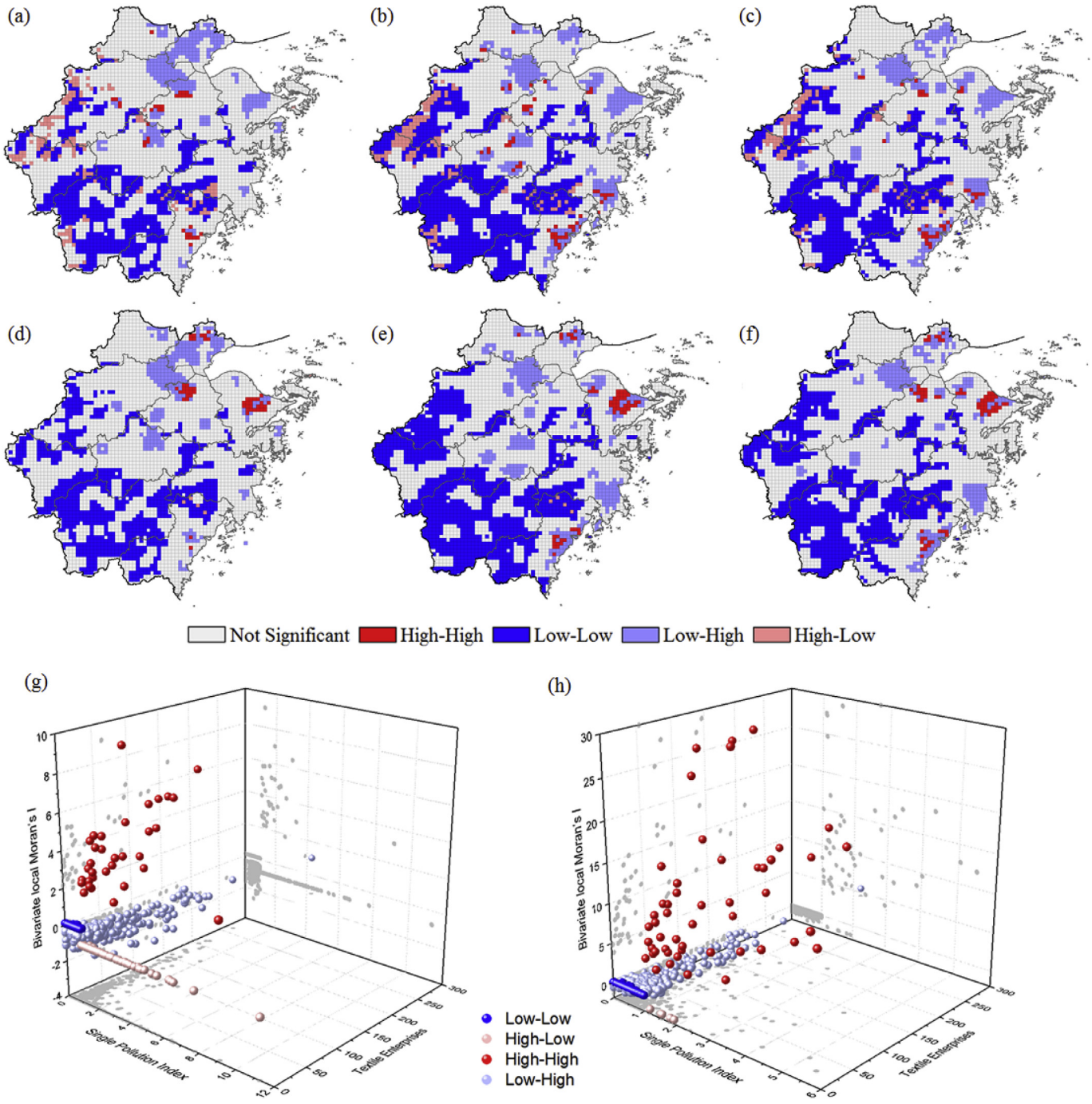

Fig. 4. Source apportionment of soil heavy metal pollution by bivariate local Moran's I model using the data of soil Cd and Hg pollution degree and pollution enterprises and 3D Scatter plots of the four scenarios of spatial correlation between heavy metals and textile industry. a: textile industry & Cd element, b: metalwork industry & Cd element, c: chemical industry & Cd element, d: textile industry & Hg element, e: metalwork industry & Hg element, f: chemical industry & Hg element, g: Cd element, h: Hg element.

environmental protection department of the Zhejiang University. Table 5 shows the derived background values of Cd and Hg elements in soils (0–20 cm). According to Fig. 4, the high-low area of Cd pollution was mainly distributed in the QZ district. The background value of Cd element in the QZ district was 0.201 mg kg^{-1}, which was obviously higher than the other districts of the study area. By summarizing the historical data in the Statistical Yearbook of the study area, the causes of the high concentration of Cd element in high-low area was speculated. In the west of the HZ district, the agriculture was relatively developed, whereas the average application of chemical fertilizer was greatly high with a value of 362 kg hm^{-2}. The excessive application of chemical fertilizer caused the serious agricultural non-point source pollution and threaten the quality and safety of agricultural products. The northern part of the QZ district had rich stone coal resource with approximately 5 billion tons, accounting for 49% of exploration reserve in the study area. In the 1970s, because of the scarce energy, stone coal was widely mined and used, resulting in the high concentration of Cd element in a later environmental background investigation of the QZ district. The excessive fertilization and coal

Table 5
Descriptive statistics for the background values of the Cd and Hg elements in soils.

Element	Statistics	HUZ	HZ	JH	JX	NIB	QZ	SX	TZ	WZ	LS
Cd	Mean (mg kg^{-1})	0.152	0.154	0.171	0.146	0.161	0.201	0.178	0.165	0.178	0.177
	CV (%)	19	20	20	18	19	17	16	22	23	18
Hg	Mean (mg kg^{-1})	0.132	0.128	0.0652	0.155	0.076	0.0788	0.0818	0.0857	0.127	0.0514
	CV (%)	35	54	28	29	28	23	36	32	39	16

HUZ, HZ, JH, JX, NIB, QZ, SX, TZ, WZ, LS were respectively the English abbreviations of the 10 provincial cities in the study area; CV: the coefficient of variation.

mining appeared to be the cause of high concentrations of Cd element in high-low area.

In general, soil organic matter (SOM) combines with metal ion mainly by complexation, and promotes adsorption of heavy metals in soils (Yin et al., 2002; Peng et al, 2018). The total amount of heavy metals in soils was measured in this study. When the adsorption in soils was strong, the total amount of soil heavy metals was large. According to Fig. 3, the concentration of SOM in the QZ and HZ districts was relatively high, and the topography of this region belonged to the basin, resulting that heavy metals were washed away difficultly by water runoff. These environmental factors caused the high concentration of Cd element in high-low area to some extent.

4. Conclusion

The performances of three classification methods were compared, which were SVM, NB, and ANN. The geographical dataset was divided into four categories, including textile industry, chemical industry, metalwork industry, and the other industry. It was found that Multinomial NB was slightly better than the other models in the performance of classifying the geographical dataset, with an accuracy of 87% and Kappa coefficient of 0.82. The high CV values of Cd and Hg elements indicated that heavy metal accumulation in soils were significantly affected by anthropogenic activities. The spatial distribution map of soil heavy metal pollution sources was calculated based on bivariate local Moran's I analysis realizations of heavy metals and enterprises. This study demonstrated that (i) Cd pollution in soils was mainly affected by excessive fertilization and coal mining and the metalwork industry had an influence on Cd pollution more seriously than the other industries, (ii) Soil Hg pollution was closely related to enterprise pollution and the chemical industry was most serious.

Competing financial interests

The authors declare no competing financial interests.

Acknowledgements

The project was planned and designed by Zhou S. and Lianqing Z.; the research data was provided by Youwei Z.; the model was constructed and analyzed by Xiaolin J.; the paper was constructed by Xiaolin J., Bifeng H. and Ben P. M.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envpol.2019.04.047.

Funding

This work was supported by the National Key Research and Development Program of China (2018YFD0800202) and Key Research and Development Project of Zhejiang Province (2015C02011).

References

Deng, M., Yang, X.X., Shi, Y., Gong, J.Y., Liu, Y., Liu, H.M., 2019. A density-based approach for detecting network-constrained clusters in spatial point events. Int. J. Geogr. Inf. Sci. 33 (3), 466–488.
Dudzik, P., Sawicka-Kapusta, K., Tybik, R., Pacwa, K., 2010. Assessment of environmental pollution by metals, sulphure dioxide and nitrogen in Wolinski National Park. Nat. Environ. Monit. 11, 37–48.
Duodu, G.O., Goonetilleke, A., Ayoko, G.A., 2017. Potential bioavailability assessment, source apportionment and ecological risk of heavy metals in the sediment of Brisbane River estuary, Australia. Mar. Pollut. Bull. 117 (1–2), 523–531.
Facchinelli, A., Sacchi, E., Mallen, L., 2001. Multivariate statistical and gis-based approach to identify heavy metal sources in soils. Environ. Pollut. 1143, 313–324.
Grazi, F., Jeroen, C.J.M., van den, Bergh, Rietveld, P., 2007. Spatial welfare economics versus ecological footprint: modeling agglomeration, externalities and trade. Environ. Resour. Econ. 38 (1), 135–153.
Guan, Q.Y., Wang, F.F., Xu, C.Q., Pan, N.H., Lin, J.K., Zhao, R., Yang, Y.Y., Lou, H.P., 2018. Source apportionment of heavy metals in agricultural soil based on PMF: a case study in Hexi Corridor, northwest China. Chemosphere 193, 189–197.
Hu, B.F., Chen, S.C., Hu, J., Xia, F., Xu, J.F., Li, Y., Shi, Z., 2017a. Application of portable XRF and VNIR sensors for rapid assessment of soil heavy metal pollution. PLoS One 12 (2), e0172438.
Hu, B.F., Jia, X.L., Hu, J., Xu, D.Y., Xia, F., Li, Y., 2017b. Assessment of heavy metal pollution and health risks in the soil-plant-human system in the Yangtze River Delta, China. Int. J. Environ. Res. Public Health 14 (9), 1–18.
Hu, B.F., Shao, S., Fu, Z.Y., Li, Y., Ni, H., Chen, S.C., Zhou, Y., Jin, B., Shi, Z., 2019. Identifying heavy metal pollution hot spots in soil-rice systems: a case study in South of Yangtze River Delta, China. Sci. Total Environ. 658, 614–625.
Hu, Y.A., 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.
Jiang, Y.X., Chao, S.H., Liu, J.W., Yang, Y., Chen, Y.J., Zhang, A.C., Cao, H.B., 2017. Source apportionment and health risk assessment of heavy metals in soil for a township in Jiangsu Province, China. Chemosphere 168, 1658.
Keerthi, S.S., Lin, C.J., 2003. Asymptotic behaviors of support vector machines with Gaussian kernel. Neural Comput. 15 (7), 1667–1689.
Khalil, E.H., 2018. Combining instance weighting and fine tuning for training naive Bayesian classifiers with scant training data. Int. Arab J. Inf. Technol. 15 (6), 1099–1106.
Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33 (1), 159–174.
Luo, X.S., Ip, C., Li, W., Tao, S., Li, X.D., 2014. Spatial-temporal variations, sources, and transport of airborne inhalable metals (PM10) in urban and rural areas of northern China. Atmos. Chem. Phys. Discuss. 14, 13133–13165.
Ma, R.X., Wang, K., Qiu, T., Sangaiah, A.K., Lin, D., Bin Liaqat, H., 2017. Feature-based compositing memory networks for aspect-based sentiment classification in social internet of things. Future Gener. Comp 92, 879–888.
Ma, W.C., Tai, L.Y., Qiao, Z., Zhong, L., Wang, Z., Fu, K.X., Chen, G.Y., 2018. Contamination source apportionment and health risk assessment of heavy metals in soil around municipal solid waste incinerator: a case study in North China. Sci. Total Environ. s631–632, 348–357.
Marchant, B.P., Saby, N.P.A., Jolivet, C.C., Arrouays, D., Lark, R.M., 2011. Spatial prediction of soil properties with copulas. Geoderma 162, 327–334.
Marchant, B.P., Saby, N.P.A., Arrouays, D., 2017. A survey of topsoil arsenic and mercury concentrations across France. Chemosphere 181, 635–644.
Piroonsup, N., Sinthupinyo, S., 2017. Analysis of training data using clustering to improve semi-supervised self-training. Knowl.-Based Syst. 143, 65–80.
Qu, M., Li, W., Zhang, C., Wang, S., Yang, Y., He, L., 2013. Source apportionment of heavy metals in soils using multivariate statistics and geostatistics. Pedosphere 23, 437–444.
Rigina, O., 2002. Environmental impact assessment of the mining and concentration activities in the Kola Peninsula, Russia by multidate remote sensing. Environ. Monit. Assess. 75 (1), 13–33.
Salles, T., Rocha, L., Mourão, F., Gonçalves, M., Viegas Jr., F.,, W. M, 2017. A two-stage machine learning approach for temporally-robust text classification. Inf. Syst. 69, 40–58.
Peng, S.M., Wang, P., Peng, L.F., Cheng, T., Sun, W.M., Shi, Z.Q., 2018. Predicting heavy metal partition equilibrium in soils: roles of soil components and binding sites. Soil Sci. Soc. Am. J. 82 (4), 839–849.
Sun, Z.J., Shangguan, Y.X., Wei, Y., Su, B.Y., Zhou, C.Z., Hou, H., 2018. A study on

antimony migration in soils using an artificial neural network model and a convection-dispersion diffusion model. Ecol. Model. 389, 1–10.
Verhoef, E.T., Nijkamp, P., 2002. Externalities in urban sustainability: environmental versus localization-type agglomeration externalities in a general spatial equilibrium model of a single-sector monocentric industrial city. Ecol. Econ. 40 (2), 157–179.
Wang, C., Yang, Z., Zhong, C., Ji, J., 2016. Temporal-spatial variation and source apportionment of soil heavy metals in the representative river-alluviation depositional system. Environ. Pollut. 216, 18–26.
Wu, S.H., Zhou, S.L., Bao, H.J., Chen, D.X., Wang, C.H., Li, B.J., Tong, G.J., Yuan, Y.J., Xu, B.G., 2019. Improving risk management by using the spatial interaction relationship of heavy metals and PAHs in urban soil. J. Hazard Mater. 364, 108–116.
Yin, Y., Impellitteri, C.A., You, S.J., Allen, H.E., 2002. The importance of organic matter distribution and extract soil: solution ratio on the desorption of heavy metals from soils. Sci. Total Environ. 287 (1–2), 107–119.
Zawadzka, M., Łukowski, M.I., 2010. The content of Zn, Cu, Cr in podzolic soils of Roztocze National Park at the line of metallurgical and sulphur and the highway. Acta Agrophys 16 (2), 459–470.