J. Geogr. Sci. 2019, 29(10): 1610-1624
DOI: https://doi.org/10.1007/s11442-019-1644-8

Exploring the database of a soil environmental survey using a geo-self-organizing map: A pilot study

LIAO Xiaoyong[1], TAO Huan[1,2,3], GONG Xuegang[1,3], LI You[1,3]

1. Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China;
2. Beijing Research Center for Information Technology in Agriculture, Beijing 100097, China;
3. University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: A model integrating geo-information and self-organizing map (SOM) for exploring the database of soil environmental surveys was established. The dataset of 5 heavy metals (As, Cd, Cr, Hg, and Pb) was built by the regular grid sampling in Hechi, Guangxi Zhuang Autonomous Region in southern China. Auxiliary datasets were collected throughout the study area to help interpret the potential causes of pollution. The main findings are as follows: (1) Soil samples of 5 elements exhibited strong variation and high skewness. High pollution risk existed in the case study area, especially Hg and Cd. (2) As and Pb had a similar topological distribution pattern, meaning they behaved similarly in the soil environment. Cr had behaviours in soil different from those of the other 4 elements. (3) From the U-matrix of SOM networks, 3 levels of SEQ were identified, and 11 high risk areas of soil heavy metal-contaminated were found throughout the study area, which were basically near rivers, factories, and ore zones. (4) The variations of contamination index (CI) followed the trend of construction land (1.353) > forestland (1.267) > cropland (1.175) > grassland (1.056), which suggest that decision makers should focus more on the problem of soil pollution surrounding industrial and mining enterprises and farmland.

Keywords: self-organizing map; geo-information; heavy metal; soil environmental quality; Hechi

1 Introduction

The heavy metal contamination of soil is one of the most pressing problems in China (Kong, 2014) and other countries (Tóth *et al.*, 2016) because of its huge recovery cost. Detailed information about the spatial distribution of the regional soil environmental quality (SEQ) and pollution risks of heavy metals can provide fundamental information for regional environmental planning and environmental management. However, it is difficult to comprehensively

Received: 2018-07-05 **Accepted:** 2018-10-31
Foundation: Strategic Priority Research Program of the Chinese Academy of Sciences, No.XDA19040302; The Key Research Program of the Chinese Academy of Sciences, No.KFZD-SW-111
Author: Liao Xiaoyong (1977–), PhD and Professor, specialized in evaluation and remediation of soil pollution. E-mail: liaoxy@igsnrr.ac.cn

assess SEQ for two reasons (Guan *et al*., 2016). One is for strong spatial variations of heavy metal content and other physical and chemical properties (Cai *et al*., 2010). This spatial variation is difficult to accurately simulate using a specific mathematical model (Bação *et al*., 2004). The other is for complex causes of soil pollution. The soil heavy metal content is affected not only by natural factors (such as parent material, climate, and topography) but also anthropogenic factors (such as soil waste accumulation, pesticide and fertilizer use), and these factors contribute to the contamination level in different ways. When the existing SEQ assessment methods were reviewed, certain classical methods, like the Nemerow pollution index (Jaffar *et al*., 2017) and the geo-accumulation index (Dotaniya *et al*., 2017), were found to be commonly used. The geostatistical approach (Pan *et al*., 2016) and sandwich model (Wang *et al*., 2013; Li *et al*., 2016) were also adopted in recent research on assessing SEQ in response to heavy metals. In most of these methods, evaluators had to specify regulated values (such as Cd of 0.2 $mg \cdot kg^{-1}$ for the Chinese standard) and threshold values (for the Nemerow pollution index, $P<1$ indicates the soil is pollution free, $1<P<2$ indicates lightly polluted soil, $2<P<3$ indicates the pollution is intermediate, and $P>3$ indicates the pollution is severe) for the evaluation indexes. Therefore, the SEQ results were significantly influenced by subjective factors.

Artificial neural networks (ANNs) are considered dependable and efficient method for solving the problems of strong spatial variations, complex causes of soil pollution, and subjectivity. ANNs can successfully address the non-linearity of systems and exhibit the features of self-learning and self-adaptiveness. ANNs began to be applied to SEQ assessment in the early 21st century to classify soil textures and soil physical properties (Cockx *et al*., 2009; Chang *et al*., 2000), predict non-point source pollution (Muleta *et al*., 2005) and soil salinization (Patel *et al*., 2007), and assess soil organic carbon (Somaratne *et al*., 2005). Back-propagation (BP) neural networks, radial basis function (RBF) neural networks and self-organizing map (SOM) neural networks are the most common ANN algorithms used in soil quality assessment. BP neural networks have been adopted to simulate the spatial distribution of soil pollutants (Li *et al*., 2011; Zhou *et al*., 2015) and to estimate heavy metal sorption and transportation (Bogusław *et al*., 2006; Anagu *et al*., 2009). RBF neural networks have also been used to assess spatial variation and contamination levels of soil heavy metals (Sakizadeh *et al*., 2016). SOM, also known as Kohonen networks, were proposed by Teuvo Kohonen (Kohonen, 1982). SOM outperforms other neural network methods in self-learning, visual interpretation, dimensionality reduction, and non-linear analysis. SOM is an unsupervised neural network algorithm (Kohonen *et al*., 2002) where the clustering results are objectively produced by the adaptive learning network without intervention. SOM neural networks enable the visualization of information and clusters of the soil heavy metal contents by mapping multi-dimensional data onto a 2-dimensional output space. Therefore, SOM has the function of compressing multi-dimensional information and preserves the most important topological and metric relationships of these data elements. The topological and metric relationships are very important to the interpretation of clustering results.

To date, a few studies on environmental quality assessment in sediment, water, or soil caused by heavy metals have combined SOM neural networks with other methods. When principal components analysis (PCA), cluster analysis (CA) and SOM were applied to a

large environmental data set of chemical indicators of river water quality, Astel *et al*. (2007) found that SOM clustering allowed the simultaneous observation of both spatial and temporal changes in water quality. Alvarez-guerra *et al*. (2008) assessed the feasibility of applying SOM for the classification of sediment quality and compared it with PCA and hierarchical cluster analysis (HCA). The results demonstrated that the powerful visualization tool of the SOM offered more information that was more easily accessible than that provided by HCA or PCA. Olawoyin *et al*. (2013) explored the capability of applying SOM neural networks to environmental quality assessment in an oil-contaminated area, focusing on the classification, interpretation and visualization of water, soil and sediment data. Rivera *et al*. (2015) applied the SOM method as an exploratory technique to interpret an unlabelled soil quality database; the findings indicated that the ability of the SOM to visualize multi-dimensional datasets provided insight into the data in the exploratory phase and provided a perspective for researchers to discover patterns from multi-dimensional data by using low-dimensional data.

However, these studies did not consider the geo-information of soil samples when using SOM. Geo-information could provide intuitive assessment results for decision makers. Researches (Yang *et al*., 2014; Huang *et al*., 2017) have indicated that deep data mining is possible when SOM and geo-information are combined. Wang *et al*. (2015) assessed the spatial characteristics of heavy metal-contaminated sediment to determine the contaminated hotspots by using SOM and factor analysis (FA). In the present study, a model for SEQ assessment is established in which the geo-information and SOM are integrated to cluster, interpret and visualize a large, high-dimensional environmental dataset of soil heavy metals. This research may provide important guidance for the assessment of SEQ and provide an approach to explore the potential causes of risk in areas with highly susceptible soils.

2 Materials and methods

2.1 Soil sampling

The study area is located in Hechi, Guangxi Zhuang Autonomous Region, in southern China, and the total sampling area is 13,464 km^2. This area belongs to the circum-Pacific polymetallic metallogenic belt with abundant mineral resources such as tin, lead, zinc, mercury, antimony, and arsenic. There are at least 100 enterprises that discharge pollutants in the sampling area, including smelting enterprises, mining and mineral processing enterprises, and chemical raw materials processing enterprises. The topography of the study area is higher in the northwest and lower in the southeast, and the elevation is between 86 and 1688 m. Regular grids were adopted to the sampling scheme with a grid size of 5 km. Considering the transportation conditions and the topography, the sampling points within each grid were selected on flat terrain and were far from major roads. At least one soil sample was collected in each grid only when the place of this grid is inaccessible. A total of 513 samples were taken from the topsoil (a depth of 0–20 cm) in the first half of 2013 (Figure 1).

All soil samples were air dried and crushed and then sub-divided into two portions. Portion one was sieved through a 20-mesh nylon screen for the analysis of soil physicochemical properties. Another portion was passed through a 200-mesh nylon sieve prior to the microwave digestion procedure to determine the contents of heavy metals. To determine the contents of

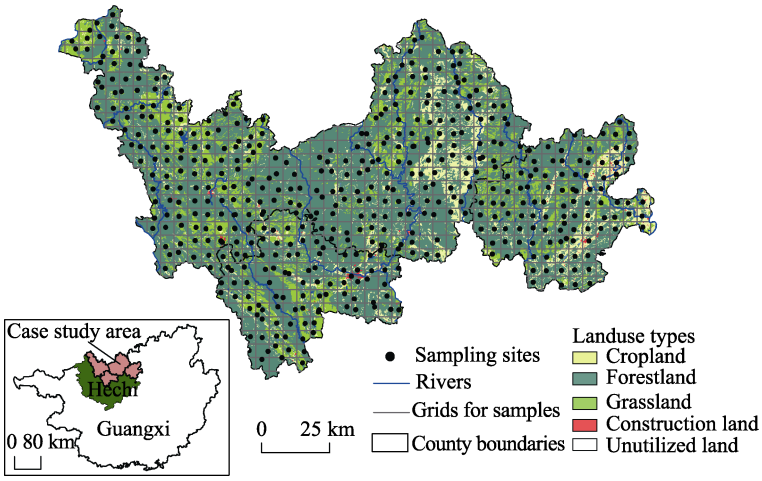

Figure 1 Soil sampling in Hechi, Guangxi Zhuang Autonomous Region in southern China

5 heavy metal elements (As, Cd, Cr, Hg and Pb) in the soils, 0.200 g of each soil sample was transferred to suitably inert polymeric microwave vessels using an acid mixture of 9 ml HNO_3 and 3 ml HF (USEPA, 3052). The content of each element was determined by atomic fluorescence spectrometry (As, Hg), graphite furnace atomic absorption spectrophotometry (Cd, Pb) and flame atomic absorption spectrophotometry (Cr). In addition, the dataset of land-use was provided by the Data Center for Resources and Environmental Sciences, Chinese Academy of Sciences (RESDC) (http://www.resdc.cn) published in 2015. The area of cropland is approximately 1631.88 km^2, which constitutes 12.20% of the study area; forestland constitutes 71.7%, grassland 16.27%, and construction land 0.47%.

2.2 Principle of the SOM network

The SOM network construction comprises an input layer and an output layer. The input layer is composed of multi-dimensional data that can be defined as a matrix $Q_{s\times p}$ with p columns (5 heavy metals) and s rows (513 soil samples). Each neuron from the output layer has a 1-dimensional, 2-dimensional or multi-dimensional grid form, and the common output is the 2-dimensional topologic structure of a regular hexagonal grid. The neighbouring neurons in the output layer are interconnected and have topologic structures. Two adjacent neurons, Y_j and Y_{j+1}, are associated with a weight vector (or codebook vector). The neurons X_i in the input layer and Y_j in the output layer have a weighted interconnection. The relationship of this interconnection in the SOM model can be described as follows (Equation (1)):

$$Y_j = f\left(\sum_{i=1}^{m} w_{ij}x_i - \theta\right) \tag{1}$$

where the weight vector w_{ij} represents the strength of these interconnections, $f(.)$ is the activation function, and θ is constant.

The second version of the SOM Toolbox for MATLAB® R2012b developed by Juha Vesanto *et al*. from the University of Helsinki (SOM tool, http://www.cis.hut.fi/ projects/ somtoolbox) was used to implement the SOMs. The SOM tool implemented an ordered dimen-

sionality-reducing mapping of the training data, i.e., the tool provided a projection of the multi-dimensional data into a 2-dimensional map preserving the topology of the input data space. The training process of the SOM tool involved a competitive learning algorithm that included 6 steps (Kohonen *et al*., 2002): 1) Variables including the weight vector $\hat{W}_j$ in the output layer, the winner constraint $N_j^*(0)$, and the learning rate $\eta(t)$ are initialized in the first step. 2) Then, input data and weight vector normalization, such as logarithmic transformation, range normalization, mean standard deviation normalization, and histogram equalization normalization, is performed. 3) A search is conducted for the winning neuron. When an input vector enters the SOM network, neurons in the output layer (or competition layer) are computed to obtain a winner neuron (or best matching unit, BMU) and then to excite it. Simultaneously, all other neurons will be unexcited. 4) All the excited neurons are updated, and the weight vector is adjusted. 5) The learning rate $\eta(t)$ and winner constraint $N_j^*(0)$ are adjusted. During the learning process, the winner constraint $N_j^*(t)$ will decrease, so does the learning rate $\eta(t)$. 6) After further learning, all the elements in the weight vector are separated from each other to represent a respective category in input space. This process continues unless the limited value meets the default value.

There are two common forms to visualize SOM networks in the SOM tool. One is the U-matrix (unified distance matrix, see Figure 4). The U-matrix displays the distance structures of different map units using a colour ramp in a 2-dimensional array of neurons, which maintains the topology and allows the identification of the clusters, boundaries and representative neurons. Clusters are map units that have smaller distances (cold colours), and borders between clusters have larger distances (warm colours). Another form is C-Planes (Component planes, see Figure 4). The C-Planes can represent both the distributions of the component values (elements of 5 heavy metals) and direct visual examination of the correlations between several component planes or the contribution to the clustering result. In the present study, every element corresponds to a component and is distinguished using different colours. If two components have a consensus pattern in space, they may be correlative or have closer pollution behaviours. The SOM tool also provides other visualization functions, such as hit maps and name labelling.

2.3 Geo-SOM procedures

As Figure 2 shows, the procedures for exploring the spatial database of the soil environmental survey consist of preliminary data statistics to describe the data quality and deep data mining combining SOM and geo-information to assess SEQ, the pollutant behaviours, and potential causes of the SEQ. Before the exploration, a spatial database of soil environmental survey was established and it included soil samples of 5 pollutants, land-use data, a distribution map of tailing ponds, a distribution map of chemical factories, the Chinese SEQ standard, and a hydrogeology map. The datasets of soil heavy metals were processed and analysed using PASW Statistics® 18.0 software, including mean, standard deviation, skewness and kurtosis. An SOM neural network was built to cluster the soil samples and then convert them into different pollutant levels. Additionally, we compared the SOM model and the results from the Nemerow pollution index based on the same set of soil samples. Within the SOM model, a vector including 5 elements constituted the input data space, and the output

was a 2-dimensional clustering result. The SEQ results could be visualized by geo-information based on the GPS coordinate of every sample; then, inferences could be made by combining with other information.

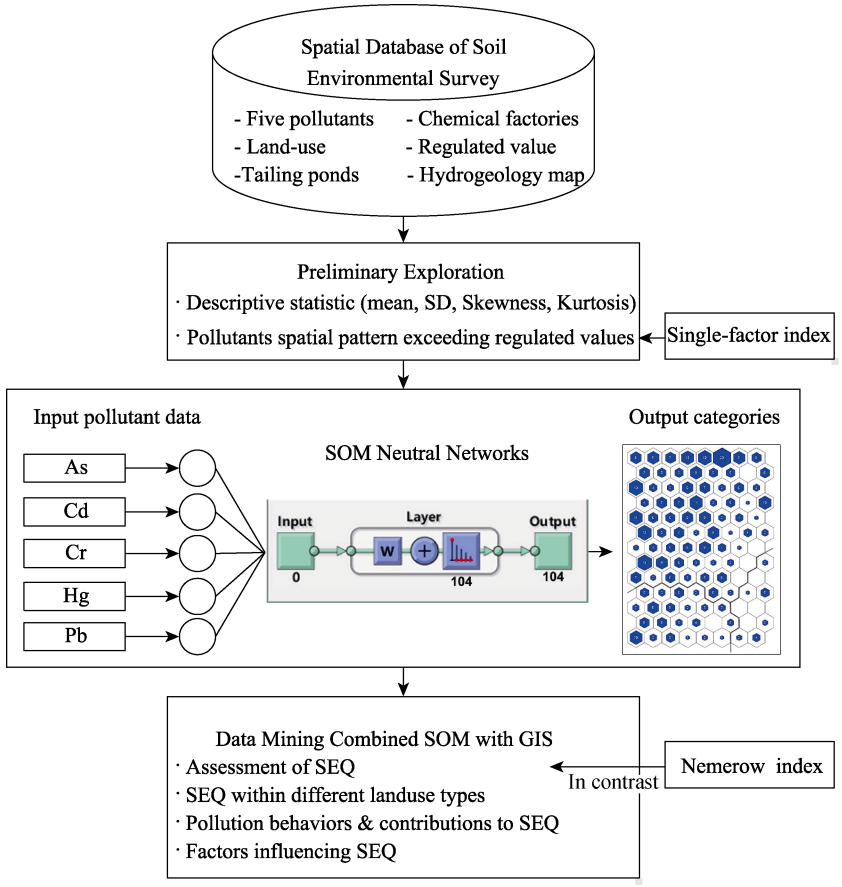

Figure 2 Flowchart of G-SOM for exploring the spatial database of the soil environmental survey

The pollution factor (P_i) quantifies the pollution of one single heavy metal, $P_i = C_i / B_i$, where C_i is the concentration of the measured pollutant, and B_i is the regulated value, which allows the levels of the different heavy metals to be determined.

The Nemerow pollution index (*PI*) (Nemerow, 1974) assesses soil quality based on the degree of pollution of various heavy metals and considering the pollution factor, as defined by the following equation (Equation 2):

$$PI = \sqrt{\frac{(\overline{P_i})^2 + (P_{i\max})^2}{2}} \tag{2}$$

where $\overline{P_i} = \frac{1}{n}\sum_{i=1}^{n} P_i$ is the mean value of the single-factor index of all elements, and $P_i$max is the maximum single-factor index of all elements. The average Nemerow pollution index ($\overline{PI}$) of each level is calculated by $\overline{PI} = \frac{1}{n}\sum_{i=1}^{n} PI_i$, where PI_i is the Nemerow pollution in-

dex of point *i*, and 'n' is the count of points at each contamination level.

The contamination index (CI) for the evaluation of land-use types on SEQ was defined as the following equation (Equation 3):

$$CI_j = \sum_{i=1}^{3} Q_i \times P_{ij} / S_j \quad (3)$$

where Q_i is the level of SEQ (i = 1, 2, 3), P_{ij} is the count of soil quality ranking i to land-use j (j = 1, 2, 3, 4), and S_j is a sum of samples with land-use j.

3 Results and discussion

3.1 Preliminary exploration

Table 1 lists the descriptive statistics of soil samples including 5 heavy metals. The contents of soil heavy metals are characterized by high variation according to their standard deviations (192.39 mg·kg^{-1} for Pb and 94.86 mg·kg^{-1} for As) and high coefficients of variation (351% for As, 280% for Cd, and 250% for Pb), which means that the elements exhibited strong variation in the sampling area. The skewness ranges from 2.25 to 16.03, and the kurtosis ranges from 9.83 to 297.07. The skewness and the kurtosis are both greater than zero, which means that the samples are high skewness. According to the Chinese standard in soil, the samples that exceeded the standard values for As, Cd, Cr, Hg and Pb among 513 samples account for 42.69%, 51.85%, 21.25%, 73.88%, and 62.00% of the total samples, respectively. This high Hg content may be related to the regional parent material and pedogenic factors.

Table 1 Sample data statistics in Hechi region

Elements	Mean (mg·kg^{-1})	Standard deviation (mg·kg^{-1})	Variation coefficient (%)	Skewness	Kurtosis	Chinese standard (mg·kg^{-1})	Exceeding rate (%)
As	27.05	94.86	351	16.03	297.07	15	42.69
Cd	1.47	4.11	280	9.15	123.09	0.2	51.85
Cr	68.27	56.86	83	2.25	9.83	90	21.25
Hg	0.48	0.72	150	5.20	39.84	0.15	73.88
Pb	76.06	192.39	250	8.50	84.73	35	62.00

The Nemerow pollution index map (Figure 3a) and single-factor index maps (Figures 3b-3f) of As, Cd, Cr, Hg, and Pb were calculated based on the regulated values of 15, 0.2, 90, 0.15, and 35 mg·kg^{-1}, respectively. The pollution scale of level 1 (PI<1), level 2 (1<PI<3), and level 3 (PI>3) was displayed in geo-information software using threshold values. A deep red colour (level 3) on the map indicates serious pollution, and a light green colour (level 1) represents a pollution free area. Overall, large areas of the sampling area shown in Figure 3a exceed the regulated value. From the perspective of a single-factor index, Hg and Cd are more serious than As, Cr and Pb, and are the main contributors to the Nemerow pollution index result. The results of the Nemerow pollution index map (Figure 3a) exhibit a similar spatial pattern to the Hg single-factor index map (Figure 3e). According to the definition of Nemerow, a defect of this method is that it will amplify the effect of heavily polluted elements, such as the Cd element in our study area, and weaken the lightly polluted elements such as Cr. Therefore, detailed pollution information will be lost in the Nemerow pollution

index map. In addition, the regulated values of the 5 heavy metals range from 0.15 to 90. Higher regulated values significantly reduce the results of the Nemerow pollution index, and in contrast, lower regulated values elevate the effects. In conclusion, the Nemerow pollution index map incorporates more information for Hg and Cd, which have relatively lower regulated values and severe pollution levels under the current grading systems, and less information for other elements.

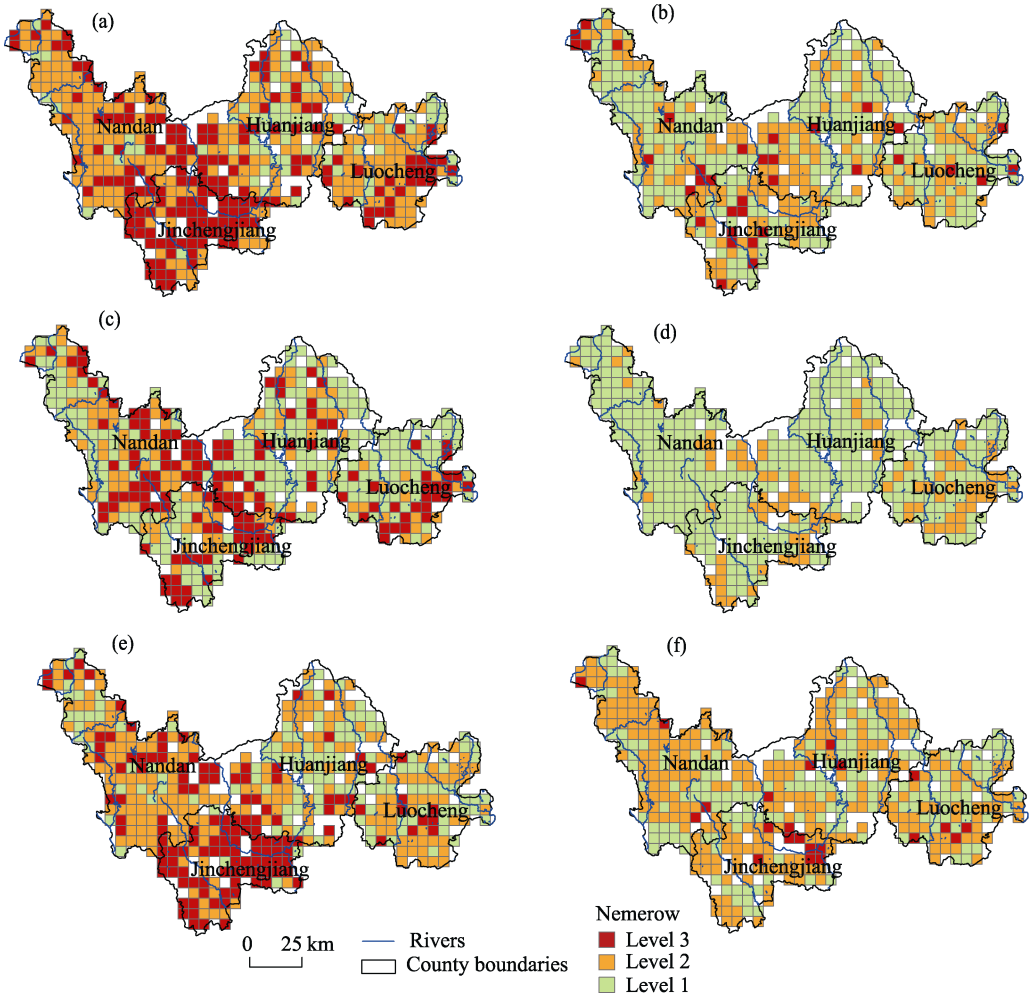

Figure 3 Spatial patterns of soil heavy metal pollution of the Nemerow pollution index map (a), and single-factor index maps of As (b), Cd (c), Cr (d), Hg (e), and Pb (f), respectively

3.2 Data mining with combined SOM and geo-information

3.2.1 Input data normalization

For the SOM tool, map size determination (the number of neurons, including its rows and columns) is an important feature. If the map is too small, it might not sufficiently represent detailed differences; if the map is too large (the number of map units is much larger than the number of samples), no new differences will be revealed, and the SOM might even be overfitted. In this study, the number of output neurons was set using Vesanto's rule (Vesanto,

2002), which defines the optimal number of neurons as $5\sqrt{s}$, where s is the number of samples. In addition, the rows and columns were evaluated using two measures as criteria – the quantization error (QE) and the topographic error (TE) – and the optimum map size was chosen based on the minimum QE and TE. Given that the training algorithm uses Euclidean distances, data must be normalized before training to avoid distortion in the results. The SOM tool provided three normalization methods: 'Var', which normalizes the variance of each variable to unity and its mean to zero; 'Range', which scales the variable values between 0 and 1 with a linear transformation; and 'Log', which is a logarithmic transformation. The results of QE and TE for the three normalization methods and different map sizes (reference to Vesanto's rule) are summarized in Table 2. The lowest values of QE (0.04) and TE (0.06) were clearly obtained using 'Range' normalization. Therefore, focusing on the results of 'Range' normalization, a 104-unit map (13×8) was selected as the best compromise between the lowest QE and TE values and with the number of neurons close to the number calculated according to Vesanto's rule.

Table 2 Summary of SOM quality measures

Rows	Columns	Map size	Variable		Log		Range	
			QE	TE	QE	TE	QE	TE
11	8	88	0.47	0.02	0.62	0.03	0.04	0.11
12	8	96	0.46	0.01	0.60	0.02	0.04	0.11
11	9	99	0.45	0.02	0.60	0.04	0.04	0.15
*13	*8	*104	0.45	0.03	0.59	0.03	*0.04	*0.06
12	9	108	0.44	0.02	0.58	0.04	0.04	0.12
11	10	110	0.44	0.03	0.58	0.04	0.04	0.18
13	9	117	0.43	0.02	0.57	0.04	0.04	0.08
12	10	120	0.43	0.02	0.56	0.04	0.04	0.14
11	11	121	0.42	0.03	0.56	0.03	0.04	0.17
13	10	130	0.42	0.02	0.55	0.03	0.03	0.14
12	11	132	0.42	0.02	0.55	0.03	0.03	0.18
13	11	143	0.41	0.02	0.54	0.03	0.03	0.15

3.2.2 SOM clustering

The U-matrix (Figure 4), as presented in the SOM output, provides the visualization of the relative distances between the neurons. Colour differential is effectively used to show the calculated distance differences between adjacent neurons. A deep blue colour on the U-matrix indicates the closeness of the vectors in the input data space, and darker red colours represent greater distances between vector values in the input data space. From the U-matrix, 3 clustering areas can be clearly distinguished, including the higher value area in the bottom right corner with darker red colours, the intermediate value area in the bottom left corner with intermediate colours, and the lower value area in the top area with deep blue colours. Nadal *et al*. (2004) used the C-Planes of the SOM networks to display the metal composition of each visual unit and found that certain pollutants may exhibit a similar behaviour. The C-Planes of As and Pb (Figure 4) display similar topological distribution patterns above the bottom right corner, which means that these elements behaved similarly in

the higher value area. A similar trend is seen in the spatial pattern of As and Pb in Figure 3. Compared with the other 4 elements, Cr has distinct differences in topological distribution patterns. The higher value area of Cr is distributed in the bottom left corner of the C-Planes, which suggests that Cr may have behaviours in soil different from those of the other 4 elements. In Figure 3, the single-factor index map of Cr shows that large areas of the sampling area are pollution free and that the spatial pattern is obviously different from those of the other 4 elements.

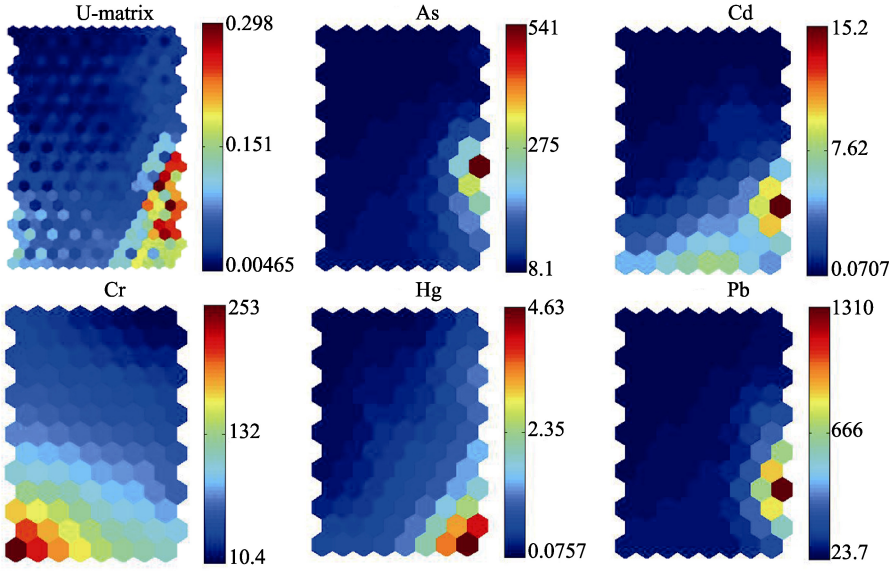

Figure 4 U-matrix of SOM and C-Planes for the five heavy metals (As, Cd, Cr, Hg, and Pb)

Similar patterns are combined with neighbouring regions in the clustering results of the SOM output network, whereas dissimilar patterns are located further apart. This process is unsupervised based on competitive learning and is thus referred to as self-organizing. In this study, the size of the dataset used in training was 513 soil sampling locations multiplied by 5 elements. After self-organized clustering, the input data were projected onto a 2-dimensional network (Figure 5) and symbolized by a colour ramp. Figure 5a shows that the entire output space is subdivided into 3 categories, including 413 samples of class 1 in the top part (lower value area), 84 samples of class 2 in the bottom left corner (intermediate value area), and 16 samples of class 3 in the bottom right corner (higher value area). Figure 5b represents the count of neuronal BMU for every category; the size of the blue regular hexagon corresponds to the number of winner neurons. The contribution of each element to the clustering results was computed based on Figure 5b and the SOM C-Planes (Figure 4). The concentrations of the 5 elements are both lower within class 1, and the contribution rates are basically not different. As the C-Planes of the SOM networks showed, Cr is characterized by the highest contribution of class 2, followed by Cd. This effect could not be observed from the Nemerow method because of its lack of a detailed description. The contribution of class 3 is mainly from Hg, Pb, and Cd, with small percentages of As and Cr. This conclusion is confirmed by the Nemerow pollution index map in Figure 3.

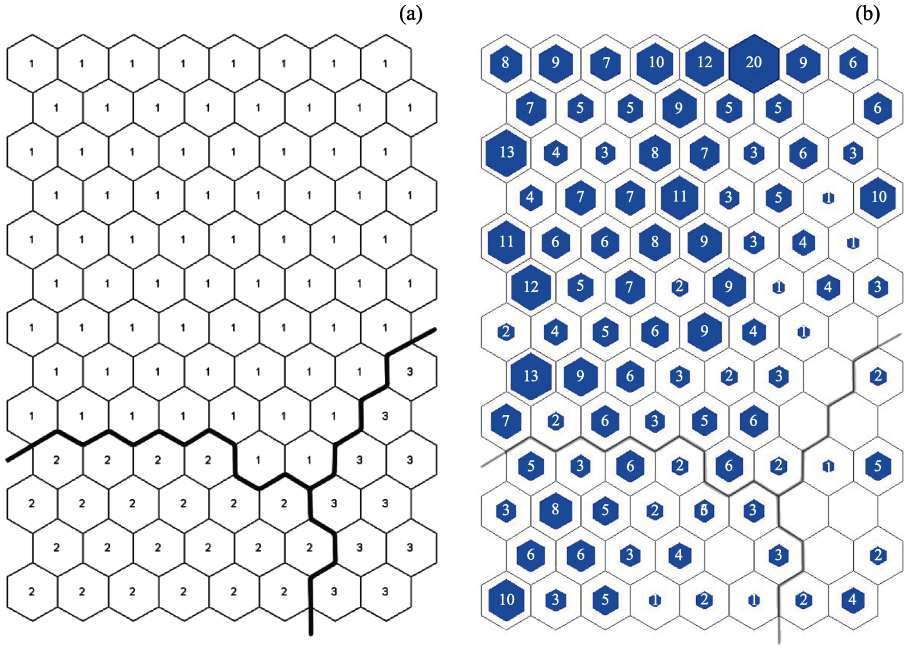

Figure 5 Results of SOM clustering of the categories map of the entire output space (a), and the map of neuronal BMU (b)

In order to test the statistical significance of the SOM clustering, spatial stratified heterogeneity q statistic, proposed by Wang Jinfeng (Wang *et al*. 2016), was adopted. The value of $q \in [0,1]$. If $q = 1$, it indicates that the result is clustering perfectly, and the small the q value is, the poor the clustering becomes. The q value of As, Cd, Cr, Hg, Pb were 0.85, 0.73, 0.76, 0.89, 0.88 respectively and the p value less than 0.05, which means the results of our clustering is statistical significance.

3.2.3 Assessment of SEQ

The SOM is capable of unsupervised competitive learning for the assessment of SEQ, and the regulated value is thus not necessary. The SOM can be used in exploring the database of soil environmental surveys when the evaluation criterion is unknown or not suitable. As Table 3 shows, class 1 (413 samples) within the lower value area is determined as level 1 (SEQ is good), class 2 (84 samples) within the intermediate value area is determined as level 2 (SEQ is medium level), and class 3 (16 samples) within the higher value area is determined as level 3 (SEQ is poor). The Nemerow pollution index of the 5 soil heavy metals for the 3 categories was also computed based on the Chinese SEQ standards in contrast with its wide application to reflect the total pollution level. Results of the Nemerow pollution index show that the mean Nemerow levels $\overline{PI}$ of class 1, class 2, and class 3 in 513 samples are 2.85, 15.18, and 49.76, respectively. This result is consistent with the SOM clustering. The same conclusion can also be drawn from the U-matrix map. The mean contents of As, Cd, Pb, and Hg correspond to the tendency that level 3 > level 2 > level 1, except that the Cr of level 2 (169.10 mg·kg^{-1}) > level 3 (86.88 mg·kg^{-1}) > level 1 (47.10 mg·kg^{-1}). This tendency could further explain why, compared with the other 4 elements, Cr exhibits different spatial patterns in soil.

Table 3 Classification statistics of SOM clusters

SEQ	Mean content of elements (mg·kg^{-1})					Counts	$\overline{PI}$
	As	Cd	Cr	Hg	Pb		
Level 1	16.74	0.46	47.1	0.33	47.24	413	2.85
Level 2	27.52	3.94	169.1	0.66	69.17	84	15.18
Level 3	290.04	8.9	86.88	3.29	782.23	16	49.76

3.2.4 Exploring causes of high-risk areas

All classified samples combined with GPS coordinate information using the spatial join tool in geo-information software were reported by a grid unit for 4 regions (Nandan, Huanjiang, Jingchengjiang, and Luocheng) (Figure 6). As Figure 6 shows, level 1 (light green) represents good soil quality or light pollution, level 3 (dark-red) represents poor soil quality or severe pollution, and level 2 (orange) is between levels 1 and 3. Overall, the SEQ is good (level 1) across most of the sampling area. The area of poor soil quality (level 3) is distributed in the surrounding rivers and mining areas.

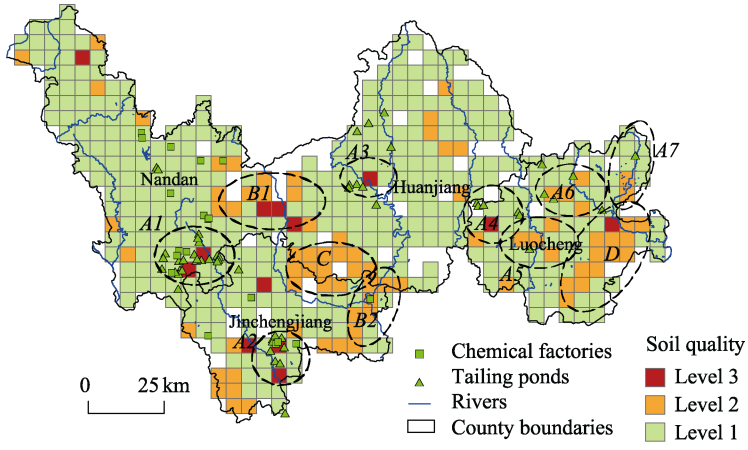

Figure 6 Spatial distribution of SEQ and 4 categories of high-risk areas of mining areas, flood areas, higher natural background value areas and sewage irrigated areas

The pollution factors of the study area can be ascribed to man-made pollution (such as wastewater irrigation, wastewater discharge, and tailing pond leakage) and natural causes (such as soil parent material, diffusion with floods, and the distribution of the mineralization belt). As Figure 6 shows, we identified a few high-risk areas which are classified into 4 categories based on the pollution factors. The districts A1, A2, A3, A4, A5, A6, and A7 are all located in mining areas and these areas have the problem of disordered piles of tailings. The study area is characterized by high annual rainfall and concentrated rainfall events, which can easily cause flash floods. During flood seasons, the pollutants deposited in river sediment discharged from factories or leaked from tailing ponds can be washed into rivers and then diffuse into the nearby soil (such as the high-risk areas of B1 and B2). Descriptions of tailing pond leakage in areas A1, A2 and A3 were obtained from the literature (Huang *et al*., 2012; Wu, 2015). These 3 regions are characterized by an abrupt change in SEQ from level 3 to level 1. Therefore, additional samples are proposed among these areas for further investigation (Gao *et al*., 2017). The high-risk area of C around the borders between

Jinchengjiang and Huanjiang is from natural pollution caused by higher natural background values affected by the weathering of parent rocks. The high-risk area of D is the low-lying farmlands and is mainly caused by sewage irrigation.

Land-use types comprehensively reflect the effects of human activities and natural environmental factors on SEQ. In the present study, the types of land-use data included cropland (194 samples), forestland (232 samples), grassland (36 samples), and construction land (51 samples). Samples of water were not considered, and water bodies were thus eliminated from the land-use data. The three levels of SEQ and the corresponding land-use types are summarized in Table 4. The variation of contamination index (CI) follows the trend of construction land (1.353) > forestland (1.267) > cropland (1.175) > grassland (1.056), which means that construction land has the worst SEQ in the study area. According to the data description of RESDC, construction land includes urban and rural construction land and industrial and mining land, among others, which suggests that decision makers should focus more on the problem of soil pollution surrounding industrial and mining enterprises. The second highest CI is for forestland, which may have been caused by the high regional soil environmental background value because our study area is in a mineralization belt. Cropland also has a problem of soil pollution that should be considered when formulating policy or environmental planning. Soil samples of level 3 account for approximately 5.9% of the total samples in construction land types. For grassland, most samples (94.4%) are at level 1, indicating a good SEQ.

Table 4 SEQ for different land-use types published in 2015

Land-use types	SEQ			Counts	CI
	Level 1	Level 2	Level 3		
Construction land	36 (70.6%)	12 (23.5%)	3 (5.9%)	51	1.353
Cropland	167 (86.1%)	20 (10.3%)	7 (3.6%)	194	1.175
Forestland	176 (75.9%)	50 (21.6%)	6 (2.6%)	232	1.267
Grassland	34 (94.4%)	2 (5.6%)	0 (0%)	36	1.056

4 Conclusions

In this study, a model integrating geo-information and SOM was established and its ability of exploring the spatial database of a soil environmental survey was examined by a case study in Hechi city, China. The result of preliminary exploration showed that soil samples of five elements exhibited strong variation and high skewness. According to the Chinese standard in soil, the samples that exceeded the standard values for As, Cd, Cr, Hg, and Pb among 513 samples accounted for 42.69%, 51.85%, 21.25%, 73.88%, and 62.00% of the total samples, respectively. The Nemerow pollution index map and single-factor index maps indicated that high pollution risk existed in the case study area, especially the Hg and Cd, which were the main contributors to the Nemerow pollution index result.

From the U-matrix of SOM networks, 3 clustering areas could be clearly distinguished, 513 samples were classified into lower value area (413 samples), intermediate value area (84 samples), and higher value area (16 samples) with the mean Nemerow levels of 2.85, 15.18, and 49.76, respectively. The C-Planes of the SOM networks showed that As and Pb had a

similar topological distribution pattern, which means that these two elements behave similarly in the soil environment. Cr had distinct differences in topological distribution patterns, which suggests that Cr may have behaviours in soil different from those of the other 4 elements.

For the deep mining, we identified a few high-risk areas (worse SEQ) and analysed their causes combined with other auxiliary information. The severe heavy metal-contaminated areas were found near rivers, factories, and ore zones. The pollution factors of the study area can be ascribed to man-made pollution (such as wastewater irrigation, wastewater discharge, and tailing pond leakage) and natural causes (such as soil parent material, diffusion with floods, and the distribution of the mineralization belt). The variation of contamination index (CI) follows the trend of construction land (1.353) > forestland (1.267) > cropland (1.175) > grassland (1.056), which suggest that decision makers should focus more on the problem of soil pollution surrounding industrial and mining enterprises and farmland.

We compared the SOM model results with results from the Nemerow pollution index. A defect of the Nemerow method is that it will amplify the effect of heavily polluted elements, especially those elements that have relatively lower regulated values and severe pollution levels under the grading systems as well as weak, lightly polluted elements. Therefore, detailed pollution information is lost in a Nemerow pollution index map. However, the SOM model can provide a U-matrix and C-Planes to visualize the process of clustering, which helps reveal the relations between different elements and their contributions to the final clustering. In addition, this model can also reduce the dependency of subjective assessment standards and grading thresholds by government or other organizations to objectively and accurately reflect the regional SEQ.

References

Alvarez-Guerra M, González-Piñuela C, Andrés A *et al.*, 2008. Assessment of self-organizing map artificial neural networks for the classification of sediment quality. *Environment International*, 34(6): 782–790.

Anagu I, Ingwersen J, Utermann J *et al.*, 2009. Estimation of heavy metal sorption in German soils using artificial neural networks. *Geoderma*, 152(1/2): 104–112.

Astel A, Tsakovski S, Barbieri P *et al.*, 2007. Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets. *Water Research*, 41(19): 4566–4578.

Bação F, Lobo V, Painho M, 2004. Geo-self-organizing map (Geo-SOM) for building and exploring homogeneous regions. In: Egenhofer M J, Freksa C, Miller H J. Third International Conference on GIScience. Berlin: Springer, 22–37.

Buszewski B, Kowalkowski T, 2006. A new model of heavy metal transport in the soil using nonlinear artificial neural networks. *Environmental Engineering Science*, 23(4): 589–595.

Cai L M, Huang L C, Zhou Y Z *et al.*, 2010. Heavy metal concentrations of agricultural soils and vegetables from Dongguan, Guangdong. *Journal of Geographical Sciences*, 20(1): 121–134.

Chang D H, Islam S, 2000. Estimation of soil physical properties using remote sensing and artificial neural network. *Remote Sensing of Environment*, 74(3): 534–544.

Cockx L, Van Meirvenne M, Verbeke L P C *et al.*, 2009. Extracting topsoil information from EM38DD sensor data using a neural network approach. *Soil Science Society of America Journal*, 73(6): 2051–2058.

Dotaniya M L, Meena V D, Rajendiran S *et al.*, 2017. Geo-accumulation indices of heavy metals in soil and groundwater of Kanpur, India under long term irrigation of tannery effluent. *Bulletin of Environmental Contamination and Toxicology*, 98(5): 706–711.

Gao B, Lu A, Pan Y *et al.*, 2017. Additional sampling layout optimization method for environmental quality grade classifications of farmland soil. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(12): 5350–5358.

Guan Y, Shao C, Gu Q *et al.*, 2016. Study of a comprehensive assessment method of the environmental quality of soil in industrial and mining gathering areas. *Stochastic Environmental Research and Risk Assessment*, 30(1): 91–102.

Huang K X, Qin L M, Wu S Z *et al.*, 2012. Situation and remedial measures for heavy metals pollution in Hechi city of Guangxi. *Journal of Guangxi Academy of Sciences*, 28(4): 320–324. (in Chinese)

Huang Y, Ye H, Zhang L *et al.*, 2017. Prediction of soil organic matter using ordinary kriging combined with the clustering of self-organizing map: A case study in Pinggu District, Beijing, China. *Soil Science*, 182(2): 52–62.

Jaffar S T A, Luo F, Ye R *et al.*, 2017. The extent of heavy metal pollution and their potential health risk in top-soils of the massively urbanized district of Shanghai. *Archives of Environmental Contamination and Toxicology*, 73(3): 362–376.

Kohonen T, 1982. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1): 59–69.

Kohonen T, Somervuo P, 2002. How to make large self-organizing maps for nonvectorial data. *Neural Networks*, 15(8/9): 945–952.

Kong X T, 2014. China must protect high-quality arable land. *Nature*, 506(7486): 7.

Li X, Gao B, Pan Y *et al.*, 2016. The soil heavy metal content mapping based on Sandwich model. In: 2016 Fifth International Conference on Agro-Geoinformatics (Agro-Geoinformatics), Piscataway: IEEE, 1–6.

Li Y, Li C K, Tao J J *et al.*, 2011. Study on spatial distribution of soil heavy metals in Huizhou City based on BP-ANN modeling and GIS. *Procedia Environmental Sciences*, 10: 1953–1960.

Muleta M K, Nicklow J W, 2005. Decision support for watershed management using evolutionary algorithms. *Journal of Water Resources Planning and Management*, 131(1): 35–44.

Nadal M, Schuhmacher M, Domingo J L, 2004. Metal pollution of soils and vegetation in an area with petrochemical industry. *Science of the Total Environment*, 321(1–3): 59–69.

Nemerow N L, 1974. Scientific Stream Pollution Analysis. New York: McGraw-Hill.

Olawoyin R, Nieto A, Grayson R L *et al.*, 2013. Application of artificial neural network (ANN)–self-organizing map (SOM) for the categorization of water, soil and sediment quality in petrochemical regions. *Expert Systems with Applications*, 40(9): 3634-3648.

Pan Y, Li H, 2016. Investigating heavy metal pollution in mining brownfield and its policy implications: A case study of the Bayan Obo rare earth mine, Inner Mongolia, China. *Environmental Management*, 57(4): 879–893.

Patel R M, Prasher S O, God P K *et al.*, 2002. Soil salinity prediction using artificial neural networks. *Journal of the American Water Resources Association*, 38(1): 91–100.

Rivera D, Sandoval M, Godoy A, 2015. Exploring soil databases: A self-organizing map approach. *Soil Use and Management*, 31(1): 121–131.

Sakizadeh M, Mirzaei R, Ghorbani H, 2017. Support vector machine and artificial neural network to model soil pollution: A case study in Semnan Province, Iran. *Neural Computing and Applications*, 28(11): 3229–3238.

Somaratne S, Seneviratne G, Coomaraswamy U, 2005. Prediction of soil organic carbon across different land-use patterns. *Soil Science Society of America Journal*, 69(5): 1580–1589.

Tóth G, Hermann T, Da Silva M R *et al.*, 2016. Heavy metals in agricultural soils of the European Union with implications for food safety. *Environment International*, 88: 299–309.

Vesanto J, 2002. Data exploration process based on the self-organizing map [D]. Helsinki: Helsinki University of Technology.

Wang J F, Haining R, Liu T J *et al.*, 2013. Sandwich estimation for multi-unit reporting on a stratified heterogeneous surface. *Environment and Planning A*, 45(10): 2515–2534.

Wang J F, Zhang T L, Fu B J, 2016. A measure of spatial stratified heterogeneity. *Ecological Indicators*, 67: 250–256.

Wang Y B, Liu C W, Wang S W, 2015. Characterization of heavy-metal-contaminated sediment by using unsupervised multivariate techniques and health risk assessment. *Ecotoxicology and Environmental Safety*, 113: 469–476.

Wu Y, 2015. Characteristics of soil heavy metal contamination around industrial and mining enterprises in Diaojiang river basin, Guangxi Zhuang Autonomous Region, China [D]. Beijing: University of Chinese Academy of Sciences. (in Chinese)

Yang C, Guo R, Wu Z *et al.*, 2014. Spatial extraction model for soil environmental quality of anomalous areas in a geographic scale. *Environmental Science and Pollution Research*, 21(4): 2697–2705.

Zhou P, Zhao Y, Zhao Z *et al*., 2015. Source mapping and determining of soil contamination by heavy metals using statistical analysis, artificial neural network, and adaptive genetic algorithm. *Journal of Environmental Chemical Engineering*, 3(4): 2569–2579.