Sustainable source analysis process of soil heavy metal pollution combining with PMF analysis method and database system

H. Jiang[1]*, Z. Wang[1], F. C. Yi[1], Q. Q. Cheng[1], C. X. Wang[1], T. Wang[1], and Y. P. Zhang[1], Q. Sun[1]

1. School of Environment and Resource, Southwest University of Science and Technology, Mianyang, China

Corresponding author: Zhe Wang (wz2004@126.com)

Key Points:

- The database platform can manage and share spatio-temporal data of soil heavy metals well.
- The process proposed in the paper can realize the continuity of research and the sharing of data.

Abstract

In this study, 27 soil samples were collected for laboratory pretreatment and the total concentration of heavy metal Cd, Cr, Cu, Ni, Pb, Zn, As and Hg was measured. Pearson correlation analysis was carried out on the measured data after removing outliers, and the comparison groups with a significant correlation at the level of 0.01 between the concentration of several groups of elements were obtained. In order to identify effectively source of soil heavy metals by PMF analysis (Positive Matrix Factorization), we drew the location map in the study area and the concentration distribution of heavy metals. Combining Pearson correlation analysis, distribution of heavy metal concentration and PMF analysis, we obtained convincing identification results of heavy metal sources. With C# language and ArcGIS Engine development components, we developed a soil heavy metal database management system to manage the spatial and attribute data needed in source apportionment for soil heavy metals, which will provide data support for the latter sustainable research. In this paper, we proposed a sustainable heavy metal pollution identification research process, SSAPD (sustainable source analysis process based on database), which includes data collection in the field, laboratory measurement, pretreatment, PMF pollution source analysis and database establishment. The process can not only effectively identify the source of soil heavy metal pollution, but also realize the continuity of research and the sharing of data.

Keyword: Soil heavy metal, PMF, Database, GIS, Source apportionment, SSAPD.

1 Introduction

At present, the hazard caused by pollution of soil heavy metals has attracted worldwide attention, including ecological stability, food security, sustainable economic development and human health and safety. Compared with developing countries, developed countries pay more attention to environmental contamination and formulate corresponding measures in the process of economic progress. Therefore, the harm degree of pollution of soil heavy metals in developed countries is relatively lower. In the developing countries, pollution of soil heavy metals in China and Russia among the most populous developing countries continues to threaten human health and sustainable economic development (Barsova et al., 2019; Huang et al., 2018; Yang et al., 2018). Based on data published in open scientific journals and environmental agency reports, Barsova et al. (2019) used a number of relevant pollution assessment indicators to assess the current situation and changes of pollution of soil heavy metals in Russia, a developing country. They found that the level of pollutants in soil near industrial enterprises remains high in most parts of Russia because soil accumulate metals in the bound species. Yang et al. (2018) assessed soil heavy metal concentration and estimated the ecological and health risks on a national scale (in China) using the data obtained throughout the document retrieval. Their results revealed that pollution of soil heavy metals and associated threats posed by cadmium (Cd), lead (Pb) and arsenic (As) are more serious. At the same time, more and more domestic and foreign scholars focus on prevention and control of contamination of soil heavy metals, including monitoring, evaluation, apportionment of source and remediation of heavy metal content in soil (Chen et al., 2019; Doabi et al., 2018; Guan et al., 2018; He et al., 2018; Jorfi et al., 2017; Khalid et al., 2017; Vareda et al., 2019).

In order to prevent and respond to its pollution of soil heavy metals effectively, apportionment of sources is essential. Only by clarifying its source can departments organically prevent and control its pollution and repair the soil polluted by heavy metals. The commonly used apportionment models of source of soil heavy metals include the Principal Component Analysis/Absolute Principal Component Scores technique (PCA/APCS), Positive Matrix

Factorization (PMF) and the method of isotope labelling method, etc. (Dong et al., 2015; Gmochowska et al., 2019; Hu et al., 2018; Mehr et al., 2017; Wang et al., 2019; Yang et al, 2020; Zhang et al. 2018). PMF model constitutes one of apportionment methods of source recommended by U.S. Environmental Protection Agency (USEPA). It mainly uses the correlation matrix and the covariance matrix to simplify high-dimensional variables and transform them into several comprehensive factors. This method does not need to measure complex original spectra, and it is an apportionment method of source with a simple and effective operation (Chen et al., 2019). By comparing PMF with PCA, Chen et al. (2019) found that the results of factor analysis of the two methods were basically consistent. Both of them were well applied to the analysis of sources of soil heavy metals, but PMF method had more advantages in clarifying the contribution rate of analytical sources. Although the above studies have done a very comprehensive analysis in terms of identifying sources of heavy metal pollution, the data are lack of management for sustainable research. The analysis of sources of soil heavy metals involves complex data, including spatial location of sampling points, field and indoor data. Moreover, when the data of multiple time periods have to be summarized, it will become more difficult to manage the data with different time and location information. In terms of spatial and attribute data involved in source analysis of soil heavy metals, its sustainability and sharing have to be extended sufficiently. At present, there are few researches on data management for source analysis of soil heavy metals at home and abroad.

In this study, the total concentration of Cd, Cr, Cu, Ni, Pb, Zn, As and Hg in soil were measured by wet digestion combined with ICP800DV instrument, and the distribution characteristics and correlation characteristics of the heavy metal concentration in soil in project region were analyzed. At the same time, this article used PMF model to identify the source of heavy metals in soil based on Pearson correlation analysis and distribution characteristics of heavy metal concentration. The purpose of this paper is to propose a sustainable research process for identification of the source of soil heavy metal pollution. The process is able to store data with integrity constraints into the database in an organized way, which will provide data support for future studies, such as identification of source of soil heavy metals and remediation of soil polluted by heavy metals under multi-temporal conditions.

2 Materials and Methods

2.1 Study area

The project region covers almost 0.37 km^2 between longitude 27°5′30″N to 27°6′5″N and latitude 102°10′2″E to 102°10′20″E in Panzhihua city, Sichuan province, China. There is a hot dry subtropical valley climate, with annual mean rainfall of 1112.6mm and an average (lowest-highest) temperature of 19.7(-2.4 to 40.3) ℃. The details of wind direction in Miyi are shown in **Table 1** and **Fig.2.** The distribution of factory, residence, road, river and project region (mainly planting mango) is shown in **Fig.1**. In this study, the soil samples were collected from the project region planted with mango trees, and there were two iron powder factories, in the west-north and due south of the project region within 3 km respectively.

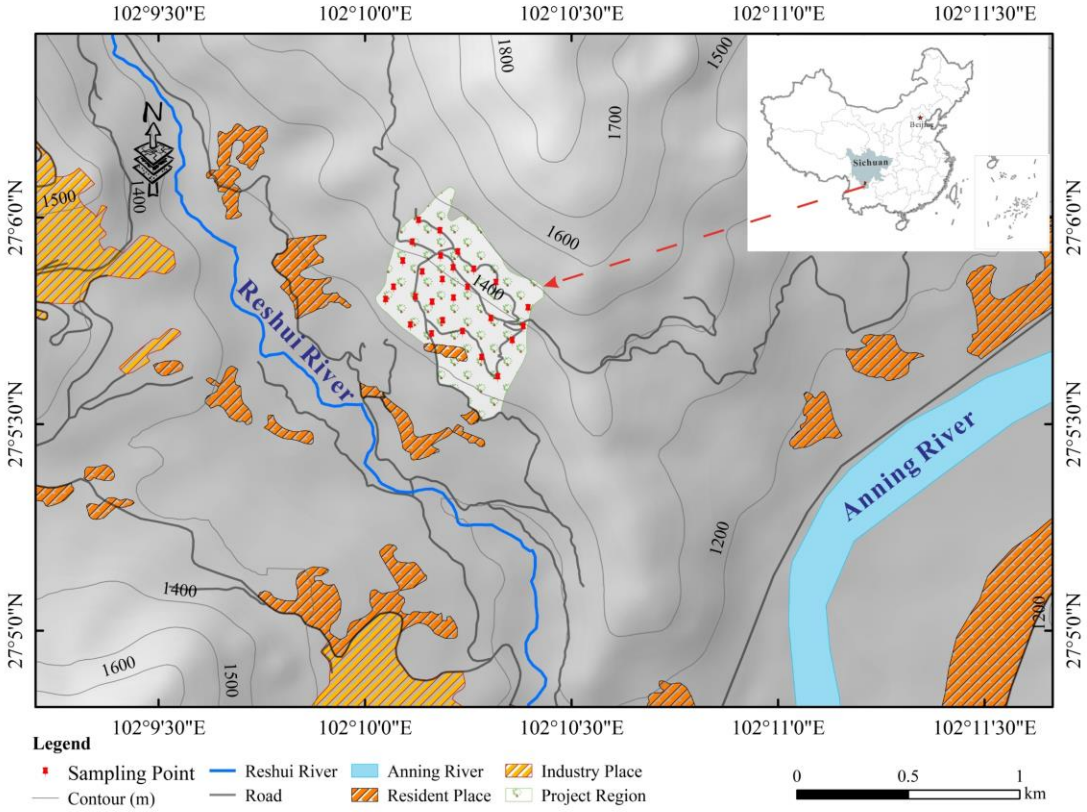


Fig 1. Map of the study area

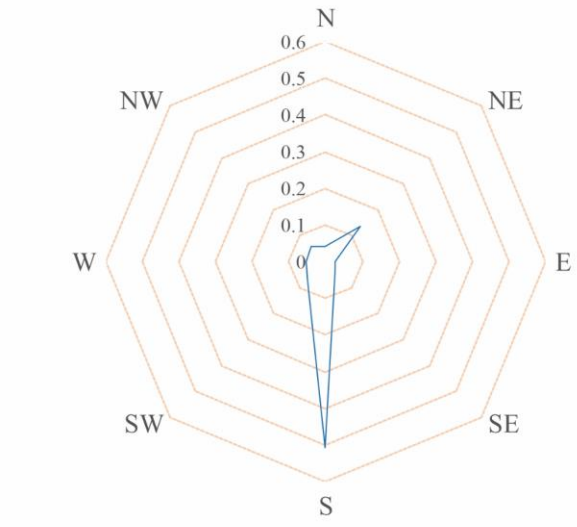


Fig 2. Rose diagram of wind direction

Table 1. Historical wind direction statistics from January 1, 2011 to December 1, 2019 in Miyi

Wind direction	N	NE	E	SE	S	SW	W	NW	No fixed wind direction

Days	128	429	84	109	1612	218	172	180	254
Frequency	0.04	0.13	0.03	0.03	0.51	0.07	0.05	0.06	0.08

2.2 Sample collection and preparation

Considering the spatial homogeneity of samples, topography in the study area, layout of factories and resident place, and the distance from the main road, we set collecting sites just like the location of sampling points in Fig.1. The land-use types mainly involve forest land (primarily planting mango trees) and industrial land. A total of 27 topsoil (0-20 cm) samples is collected in the project region, and the coordinates of each sample is recorded by GPS (Global Positioning System). Being a composite of five subsamples within 100 m^2, each soil sample is collected in a sealed polyethylene bag (Zhang et al., 2018).

The soil is air-dried at room temperature, homogenized, and ground to pass through a 0.15 mm mesh nylon sieve for measuring soil pH and total heavy metal. In the process sifting the soil, the tools, especially the mesh nylon sieve, used in this process are dried in an air-circulating oven at 50 ℃ and reused again after being thoroughly washed and rinsed with deionized water. As general characteristics, soil pH is measured by tester of pH in 1:2.5 soil to water ratio mixture (Niu et al., 2013).

2.3 Sample digestion extraction

Total heavy metal in soil is digested by high-pressure digestion. 0.15g of 0.15 mm sieved soil is put into a Teflon tank with 5mL HNO_3(ρHNO_3 = 1.42g/mL) for removal of organic matter in half an hour, then mixed by 2 mL H_2O_2(ρH_2O_2 = 1.49g/mL) and 2mL HF(ρHF= 1.49g/mL), covered and tighten with a steel-less sleeve, put into a constant temperature drying box at 180℃ for 4 h, cooled to room temperature, moved to a colorimetric tube, and filled with 2% nitric acid solution until volume to 25 mL for further analysis.

2.4 Heavy metal concentration analysis

Total heavy metal in digested solution is measured using inductively coupled plasma emission spectrometry by an Inductively Coupled Plasma-Atomic Emission Spectrometer (ICP800DV, TMO, USA) in Southwest University of Science and Technology Analysis and Testing Center in Sichuan province, China. The detection limit of the instrument is generally from tens ug/L to hundreds ug/mL.

For quality control, reagent blank and an external reference material are integrated into each batch of digestion and analysis. Precision of replicating analysis is more than 95%. The risk control standard for soil contamination of agricultural land we used is the GB 15618-2018 which is issued by the State General Administration of Market Supervision and administration, Ministry of Ecological Environment, in China.

2.5 Descriptive and spatial statistics

Descriptive statistics on total concentration of soil heavy metals are obtained. The correlation among concentration of heavy metals in soil and soil properties is determined. Log or None transformed heavy metal data obeys normal distribution by the Kolmogorov-Smirnov test. Spatial distribution of total heavy metal is determined through IDW (Inverse Distance Weighting) by using geostatistical analyst module of ArcGIS. Furthermore, we adopt the validation method

'cross validation' to compare the simulating result and the corresponding measurement concentration of total heavy metal. The validation results are detailed in 3.2 below.

3 Results and analysis

3.1 Concentration of heavy metals in soil

The soil pH in project region ranges from 4.886 to 8.068(mean = 5.954, n=27) with a coefficient of variation (CV) 13%, which show that this zone is almost in a weak acidic environment. The soil samples exhibit a wide range of Cd concentration, from 0.684 to 6.079 mg/kg (mean = 3.635 mg/kg, median = 3.241 mg/kg, n=27), with a CV (38%). More details were showed in Table 2.

Table 2 Descriptive statistics of total heavy metal concentration in soils and coefficient of variation (CV) of 27 specimens in project region

Concentration (mg/kg)					CV (%)
	Min(mg/kg)	Max(mg/kg)	Median(mg/kg)	Mean(mg/kg)	
Soil total Cd	0.684	6.079	3.241	3.635	38
Soil total Cr	30.416	577.828	79.972	132.062	86
Soil total Cu	16.963	242.918	34.256	67.347	90
Soil total Ni	9.259	211.184	38.305	57.353	74
Soil total Pb	13.918	59.361	26.041	26.861	36
Soil total Zn	10.274	121.253	69.547	63.726	38

Total Cd in all soil samples exceed the most generous critical value of Cd concentration in the risk control standard for soil contamination of agricultural land (GB 15618–2018) (0.6 mg/kg, pH>7.5). Total Pb is much lower than the most stringent critical value of Pb concentration in (GB 15618-2018) (70 mg/kg, pH $\leqslant$ 5.5), as well as total Zn (200 mg/kg, pH $\leq$ 5.5 for GB 15618-2018). Total concentration of Cr, Cu and Ni, at least one, is higher than the most generous critical value of its concentration in (GB 15618-2018), such as Cr(150 mg/kg, pH $\leq$ 5.5), Cu (150 mg/kg, pH $\leq$ 5.5) and Ni (60 mg/kg, pH $\leq$ 5.5).

There are no statistically significant correlations between pH and concentration of total Cd, Cr, Cu, Ni, Pb, Zn, respectively. Total Cd in soil shows a significant positive correlation with Cr (R=0.763**), Cu (R=0.779**), Ni (R=0.803**) and Zn (R=0.827**) at the 0.01 level. Total Cr in soil also represents a positive correlation with Cu (R=0.581**), Ni (R=0.956**) and Zn (R=0.559**) at the 0.01 level. Total Cu in soil still exhibits a positive correlation with Ni (R=0.661**) and Zn (R=0.742**) at the 0.01 level. There's a positive correlation between concentration of total Ni and Zn (R=0.599**), at the 0.01 level. More information is presented in Table 3.

Table 3 Pearson correlation coefficient (PCC or R) between concentration of total Cd, Cr, Cu, Ni, Pb, Zn and the soil pH respectively

PCC or R						
	pH	Cd	Cr	Cu	Ni	Pb	Zn
pH	1						
Cd	0.321	1					
Cr	0.295	0.763**	1				

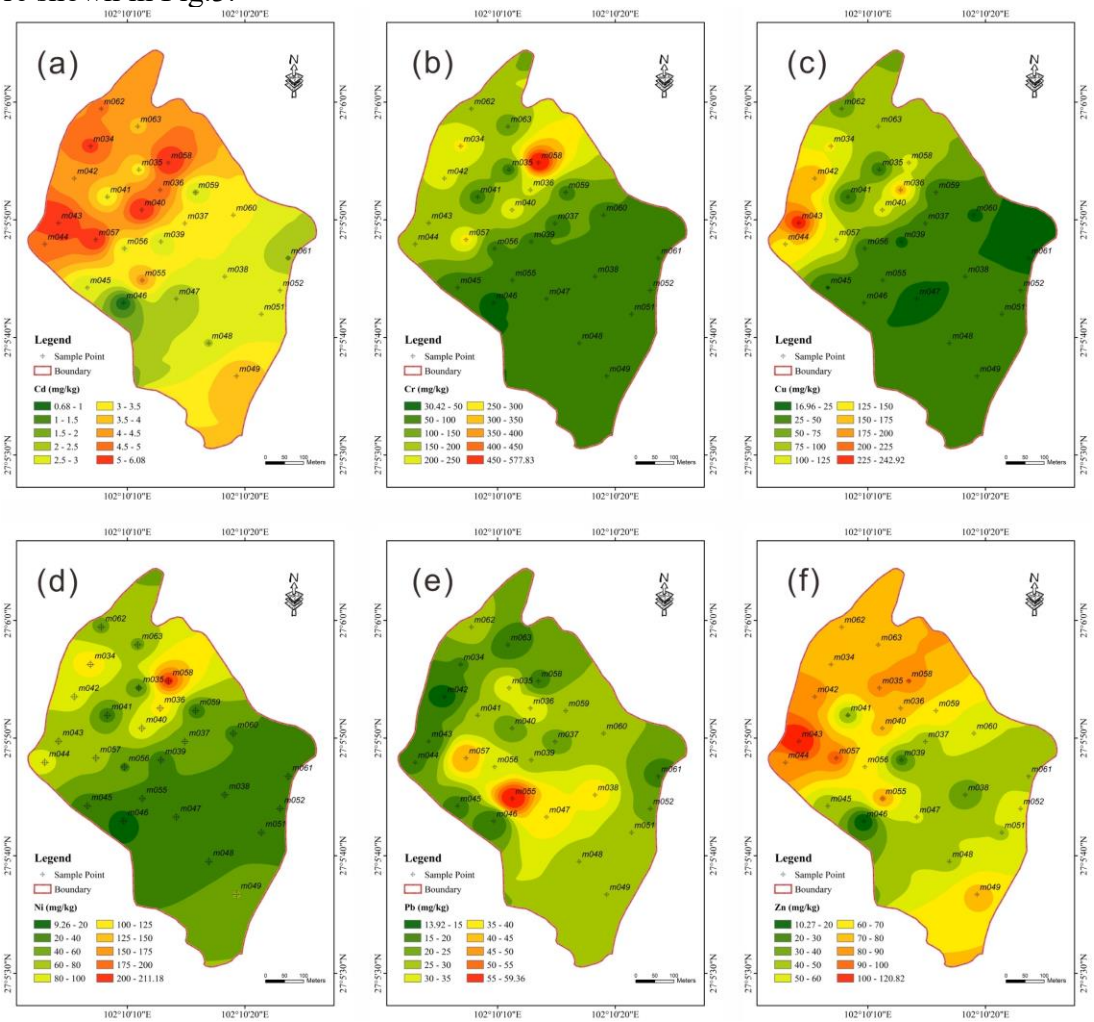
Cu	0.401*	0.779**	0.581**	1			
Ni	0.343	0.803**	0.956**	0.661**	1		
Pb	-0.075	0.035	-0.165	-0.274	-0.213	1	
Zn	0.344	0.827**	0.559**	0.742**	0.599**	0.074	1

169 * Correlation is significant at the 0.05 level.

170 ** Correlation is significant at the 0.01 level.

171 3.2 Topsoil heavy metal spatial distribution

172 The results of validation of simulating total heavy metal concentration in ArcGIS showed
173 that the interpolation results are meaningful with the Mean Standardized closed to 0 and the Root-
174 Mean-Square standardized prediction error approaching to 1. Based on results of interpolation, we
175 found total Cd in soil exhibited lower concentration in the center of the project region where total
176 Pb represented higher inversely, and radially increased towards the periphery, with significant Cd
177 concentration zones in the northwest and non-significant in the south, similar to the spatial
178 distribution of total Cr, Cu, Ni and Zn. More details of distribution of total Cd, Cr, Cu, Ni, Pb and
179 Zn were shown in Fig.3.



180

Fig 3. Concentration distribution of soil heavy metals in the project region. (a) Total Cd concentration (mg/kg), (b) Total Cr concentration (mg/kg), (c) Total Cu concentration (mg/kg), (d) Total Ni concentration (mg/kg), (e) Total Pb concentration (mg/kg), (f) Total Zn concentration (mg/kg)

3.3 PMF analysis

The PMF that the U.S. Environmental Protection Agency has recommended for source apportionment was developed based on Paatero and Tapper (1993,1994) and Paatero (1997). Source profiles matrix F_{ik} and source contributions matrix G_{kj} of pollution can be calculated by the PMF model (Guan et al., 2018), as shown in Equations (3-1).

$$X_{ij} = \sum_{k=1}^{p}(F_{ik}G_{kj}) + D_{ij} \qquad (i = 1,2,\ ,N; j = 1,2,\ ,M) \tag{3-1}$$

Where i , j is the number of samples and chemical species respectively, and X_{ij} is the concentration of the jth heavy metal pollutant in the ith sample, and F_{ik} is the contribution of the kth source in the ith sample, and G_{kj} is the contribution of concentration of the jth pollutant in the kth source, and the D_{ij} represents the random error.

Factor profiles and factor contributions are determined by the PMF model minimizing the objective function Q (Equation (3-2)).

$$Q = \sum_{i=1}^{N}\sum_{j=1}^{M}\left[\frac{x_{ij}-\sum_{k=1}^{p}F_{ik}G_{kj}}{u_{ij}}\right]^2 \tag{3-2}$$

For PMF analysis, two input files are essential, including one that records the concentration of the sample species and another one that records the uncertainty. The uncertainty of concentration is calculated as following Equation (3-3 and 3-4).

$$For\ c\ \leq MDL, u_{ij} = 5/6 \times MDL \tag{3-3}$$

$$For\ c\ > MDL, u_{ij} = \sqrt{(error\ fraction\ \times c)^2 + MDL^2} \tag{3-4}$$

Where c and MDL are the concentration of the chemical species and the species-specific method detection limit respectively, and error fraction is a percentage of the measurement uncertainty (Guan et al., 2018).

3.3.1 Removing outliers using the scatter plot

In order to study whether outliers could affect the results of PMF analysis, Wei et al. (2018) conducted PMF analysis with the data with and without outliers respectively. They found that, by comparing the results with eliminating outliers and not, the composition spectrum and the number of pollution sources was both the same, but the total contribution rate of each pollution source changed. The outliers must be removed in order to effectively identify the sources of pollutants.

Scatter plot is a commonly used outlier test method. Through Pearson correlation analysis, two significantly correlated elements at the level of 0.01 are respectively used as X and Y axes to draw a scatter plot. The sample number of outliers was determined by the scatter plot, and outliers were removed.

According to the correlation analysis results in Table 3, two elements with high correlation between elements were selected as X and Y axes, respectively, to draw a scatter diagram. There were two outliers, and names of the sampling point were M43 and M58, respectively. Two typical scatter diagrams were selected, as shown in Fig.4. The points circled in the black box are outliers.

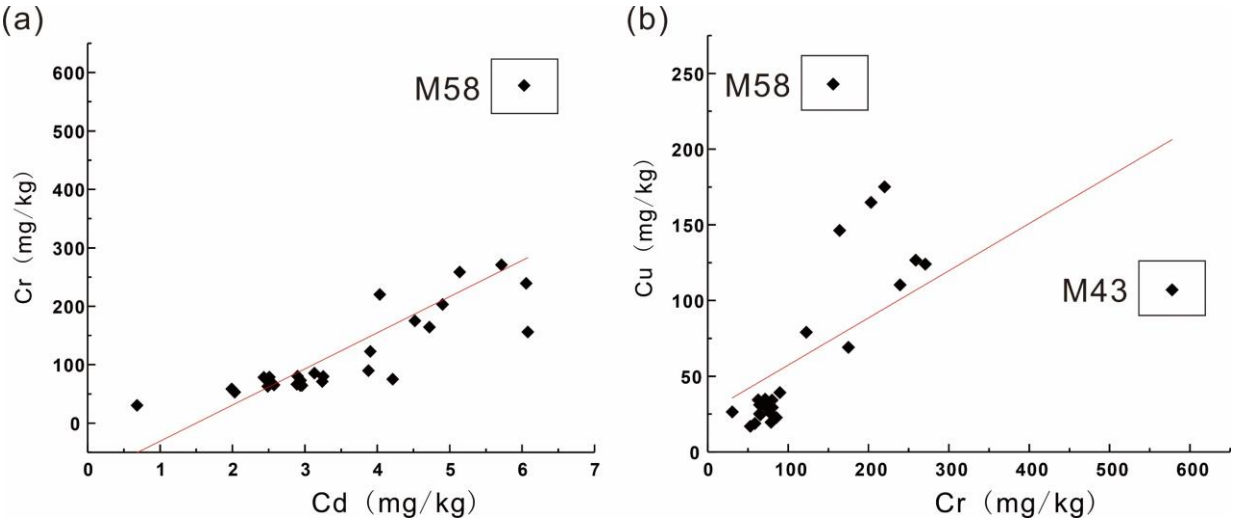


Fig 4. Scatter diagram. (a) The concentration of Cd is the horizontal axis, and the concentration of Cr is the vertical axis. (b) The concentration of Cr is the horizontal axis, and the concentration of Cu is the vertical axis.

3.3.2 Analysis of pollution source of heavy metals using PMF model

In general, PMF model requires the preparation of two types of data, including concentration and uncertainty. Before using PMF 5.0 for pollution source analysis, the signal-to-noise ratio (S/N) was calculated based on concentration and uncertainty.

Eberlv S (2005) indicated in the user manual that the PMF model calculated the signal-to-noise ratio (S/N) according to the concentration and uncertainty. After initially loading the element data, the signal-to-noise ratio (S/N) was greater than 2, which was defined as "strong". In order to get reasonable results, the user needs to debug the "strong", "weak" and the number of factors of the element several times after running the model, so that the fitting coefficient of the measured/simulated concentration of the element is greater than 0.5, and the difference between the Q value and the theoretical Q value is less than 10%.

The source contribution rate of heavy metal elements obtained through PMF model analysis is shown in Table 4. Jia et al. (2013) found that Zn in some living areas in Zhangzhou city was significantly higher than that in other areas due to the large flow of people and garbage accumulation. On site, we found that there was more than one garbage accumulation site in the project region, which was used to store household garbage and rotting and substandard mangoes. The contribution rate of source 1 to Zn was as high as 41.47% and higher than the other four sources. Therefore, it is speculated that source 1 is living source.

By using PMF 5.0 analysis, we can directly understand that Pb is mainly affected by pollution source 4 and source 2. Chen et al. (2019) found that one of the sources of Pb was emission from motor vehicle exhaust and automobile wear. Because the project region is located in the countryside, there is not much traffic except during the mango harvest season. Total concentration of Pb in all soil samples is much lower than the most stringent critical value of Pb concentration (GB 15618-2018) (70 mg/kg, pH $\leqslant$ 5.5). Combined with the concentration distribution in the project region, the concentration of Pb decreasing to the surrounding areas shows around the only housing area in the project region. The distribution of the concentration of soil heavy metals is

consistent with the phenomenon that there is more traffic in the vicinity of residential buildings than in the outer surrounding areas. The concentration distribution of Pb is quite different from that of Cd, Cr, Cu, Ni and Zn, as shown in Fig. 3. The contribution rate of source 2 to Pb is as high as 32.09%, while the contribution rate to Cd, Cu, Ni and Zn was all less than 6.5% (except that Cr was 19.24%). Therefore, it can be pointed out explicitly that source 2 is the traffic source.

Mineral resources are rich in Pan-Xi region, and the smelting amount of ore is huge with up to 2926944 tons of smelting slag at the end of 1997. In the smelting process of zinc ore, copper ore and other ores, heavy metals such as Cd, Cr, Cu, Ni and Zn entered the environment and accumulated due to improper treatment of early smelting waste residue (Xu et al, 2000). From the map of the study area shown in Fig.1, we can find that there are two factories that make fine iron powder within 3 km near the project region. The larger factory is located in the northwest with a total area up to 0.5 km^2. The raw material of these two factories is just right ore. The production activities of the them will lead to increasing contents of certain heavy metal. PMF analysis showed that, except Pb, source 3 had a certain contribution rate to Cd, Cr, Cu, Ni and Zn, and to Cu was as high as 54.33%. Combined with the concentration distribution of soil heavy metals shown in Fig.3, we found that the contents of soil heavy metals (including Cd, Cr, Cu, Ni and Zn) decreased with increasing distance from the bigger iron powder plant's industrial, central area. Therefore, it inferred that the source 3 is the industrial source.

The contribution rate of source 4 to Cd, Pb and Zn was greater than 25%. It can be seen from table 1 that the variation coefficients of these three elements are smaller than those of other metal elements, indicating that the spatial variation degree of these three elements is small, and they are not significantly affected by human activities, like the ideas in earlier studies (Wei et al., 2018). Therefore, it is inferred that source 4 is the soil parent material and belongs to the natural source.

Due to its unique southern subtropical climate, Pan-Xi region is one of the famous hot crop areas in China. It has the greatest potential for agricultural resources' development in Sichuan province and the most prominent characteristics of agricultural products. Miyi county is the national key agricultural base during the period between 1991 and 1995. Cd, Cu, Cr and Pb in phosphate fertilizer entered the soil and accumulate when farmers applied phosphate fertilizer to crops in pursuit of high yield (Chen et al., 2009; Lu and Shi, 1992; Wu and Wang, 2000; Wang, 2014; Zhou et al., 2008). But under normal planting activities, the effect of phosphate fertilizer is limited and localized when applying on the accumulation of trace elements in the receiving soil (Jiao et al., 2012). Uprety et al. (2009) analyzed the effects of long-term application of organic and inorganic fertilizers on total concentration of trace elements of heavy metals in cultivated land on a permanent arable field exiting more than 50 years, and found that total concentration of As, Cd and Cr were highest in single superphosphate, those of Cu, Mn and Ni were highest in poultry litter and those of Pb and Zn were highest in dung water. According to information collected from field trips, nearby residents are located at the lower elevation in the project region which requires a large area of fertilization, which makes it more difficult for human to fertilize with dung water, which is in accordance with the PMF analysis results that the contribution rate of source 5 to Pb and Zn is small. In order to improve the quality of planting products, farmers had to use phosphate fertilizers and organic fertilizers (such as poultry litter) with low water content and easy to carry on, and induced Cd, Cr, Cu and Ni enter the soil, which is just in accordance with the high contribution rate of source 5 to Cd, Cr, Cu and Ni. In summary, source 5 is the agricultural source. Based on the risk control standard for soil contamination of agricultural land (GB 15618–2018),

the exceeding rate of Cd in the project region is much higher than that of other heavy metal. However, the contribution rate of the source 5 to Cd is less than that of Cr, Cu and Ni, which indicates that source 5 is not the main reason for the exceeding of Cd.

Table 4 Source contribution rate of elements of the sample set by PMF

Element	Source contribution rate/%				
	Source 1	Source 2	Source 3	Source 4	Source 5
Cd	25.99	6.42	13.46	30.25	23.87
Cr	19.15	19.24	15.39	<0.01	46.22
Cu	<0.01	<0.01	54.33	2.80	42.86
Ni	23.31	<0.01	17.91	22.00	36.78
Pb	18.34	32.09	<0.01	47.72	1.85
Zn	41.47	4.29	28.92	25.32	<0.01

The linear fitting correlation coefficient r^2 of Cr, Cu, Pb and Zn models obtained through PMF analysis was greater than 0.99, showing a great correlation (except that r^2 of Ni and Cd was 0.96 and 0.93 respectively).

3.4 Establishment of soil heavy metal database platform

In order to conduct sustainable research in this project region for future multi-temporal research, we built a database management platform for soil heavy metals by using the object-oriented advanced computer programming language C#. The database platform using structured query language SQL conduct attribute query and spatial query on the sampling points. Users can obtain the spatial and attribute information of the sampling points efficiently, and realize the connection and communication with the database.

3.4.1 Content of the database

The soil heavy metal database manages the spatial, attribute and multimedia data of the sampling points.

1) Spatial data includes remote sensing images, building graphics, road data and other relevant data of the project region, remote sensing image data such as remote sensing digital orthophoto image, and graphic data such as sampling points, factory and residential distribution of the study area.

2) The attribute data related to spatial elements is a sheet which reflects the characteristics of associated elements in a particular way, mainly including the concentration data of heavy metals measured in laboratory, soil property data, and some other information related to pollution index. Researchers can extract the data needed by the research through this platform from these property sheets and build the corresponding data table.

3) Multimedia data is an important part of soil heavy metal data, including images taken during the collection of soil samples, hand-drawn sampling site plans and sections.

The structure of soil heavy metal database is shown in Fig.5 (a).

3.4.2 Structure design of the database

1) Conception framework design

The main work in the conception framework design stage is to abstract the user requirements obtained from the requirement analysis into the information structure diagram, which is the key of the whole database design. E-R diagram of conception framework design of soil heavy metal database is shown in Fig.5 (b):

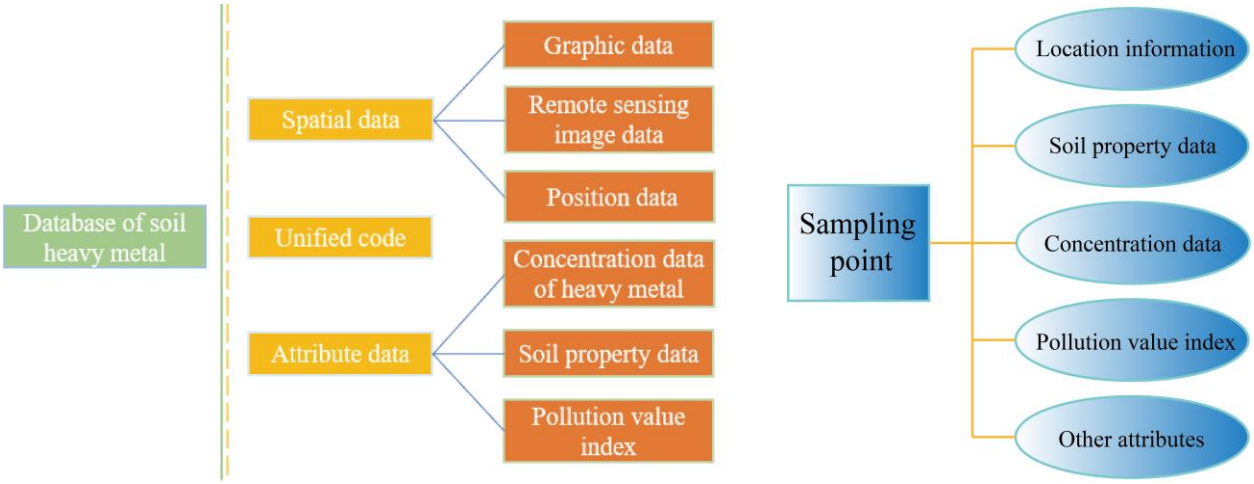


Fig 5. Structure diagram. (a) Structure of database. (b) Structure of conceptual design, E-R

2) Logic framework design of attribute data

The task of this stage is to transform the conceptual structure into a relational model and realize the logical framework design of the data. The sampling point objects determined in the conception framework design and their soil property data, heavy metal concentration data, pollution index data and location information were converted into a relational model and assigned attributes.

3) Physical framework design

The data logic model of soil heavy metal database adopts a three-level hierarchical logical structure, including a database partition, element data set and element data. All the data of the same spatial reference are stored in the same spatial database. ArcCatalog equipped with ArcGIS is used to manage the spatial data of the sampling points. The attribute information of the sampling points is stored in the relational database, and the relational data are used to establish it.

3.4.3 Establishment of the data table

Before the establishment of the database, the data should be classified and processed, and the attribute data describing a characteristic of the sampling point should be classified into one kind to make the attribute data clear and orderly, and prepare for the entry of the attribute data. Considering that the soil heavy metal database will store different temporal data to provide data support for the later multi-temporal analysis, the attribute database needs to add temporal attributes for each record to be stored. A two-dimensional property table is constructed with the basic information of sampling points and stored in the database, as shown in Table 5.

Tabel 5 Basic information of sampling points

Field Name	Description	Data type	Length
Code	The coding method used in field sampling	String	254
ID	The encoding method used for data entry	String	50
Province	The province where the sampling point is located	String	254
County	The county where the sampling point is located	String	254
Study region	Name of sampling area	String	50
Longitude (°)	Longitude in decimal	Double	15
Latitude (°)	Latitude in decimal	Double	15
Cd	Total measured value of heavy metal Cd	Double	15
Ni	Total measured value of heavy metal Ni	Double	15
Cr	Total measured value of heavy metal Cr	Double	15
Sb	Total measured value of heavy metal Sb	Double	15
As	Total measured value of heavy metal As	Double	15
Hg	Total measured value of heavy metal Hg	Double	15
Zn	Total measured value of heavy metal Zn	Double	15
Comprehensive Nemero index of pollution	Nemero pollution index is one of the indexes that express the comprehensive pollution of heavy metal at home and abroad.	Double	15
Soil property	Soil type at sampling site	String	254
Time	Field sampling time	Date Time	8

353 3.4.4 Development of a database management system

354 **1）Main interface design**

355 The main interface of the system is divided into five parts, and the top is the menu bar,
356 through which all the functional modules of the system can be directly accessed. The middle part
357 is the workspace, which is used to respond to the user's operation on the spatial data and the display
358 of the spatial data, such as highlighting the sample points retrieved by the query function. On the
359 left side of the workspace is the layer controller, through which you can set the selected layer not
360 to be displayed or even delete it. The right side of the workspace is the legend display column. The
361 main interface of the system is shown in Fig 6.

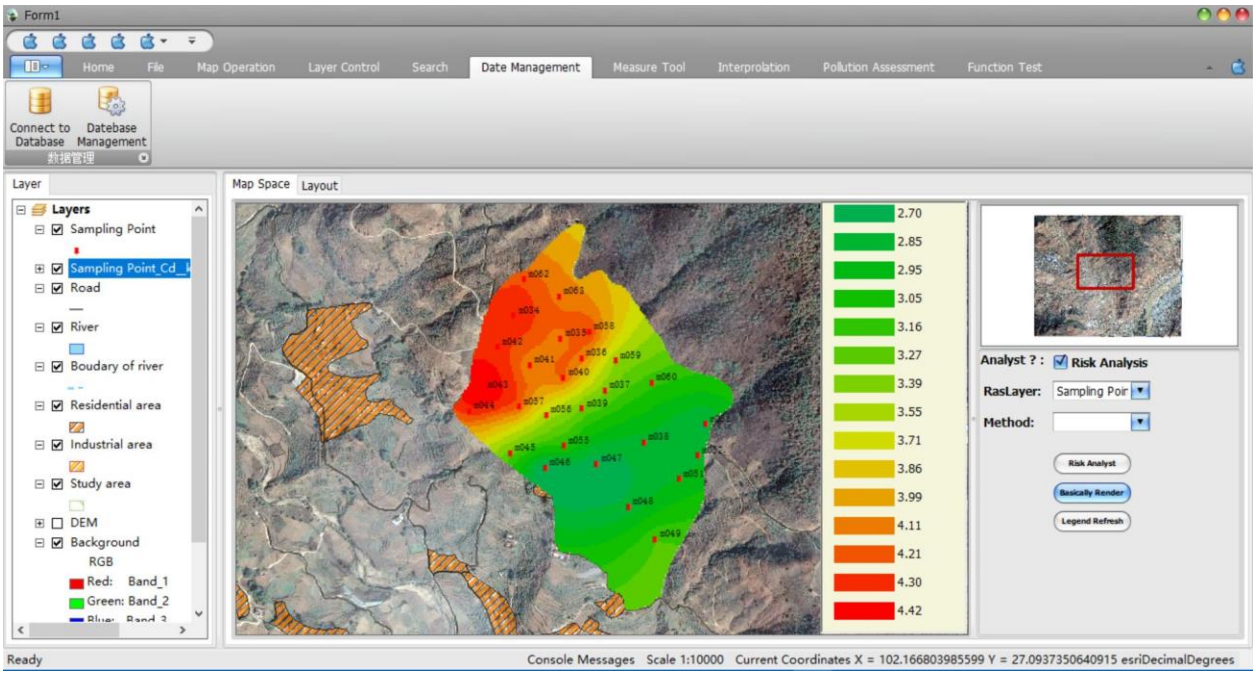


Fig 6. System main interface

2）System architecture design

In order to improve the efficiency of data editing, retrieval and analysis, the system adopts the method of joint analysis and processing of spatial and attribute data. ArcGIS Engine provides object library encapsulated according to the COM standard, supports cross-platform development, and realizes advanced analysis functions such as topology and network (Han et al., 2008). The system is divided into 7 modules based on ArcGIS Engine technology, including data collection and quality control, database management, query statistics, spatial analysis, data product services and thematic mapping. To enhance the independence of the parameter transfer, the system is divided into relatively independent functional components that communicate with each other based on interfaces. The software adopts the C/S structure, namely Client/Server structure. The specific structure system consists of data layer, GIS processing layer and service layer, which are respectively responsible for realizing data management access, GIS task processing, user interaction and other functions to improve the flexibility and maintainability of the system, as shown in Fig.7.

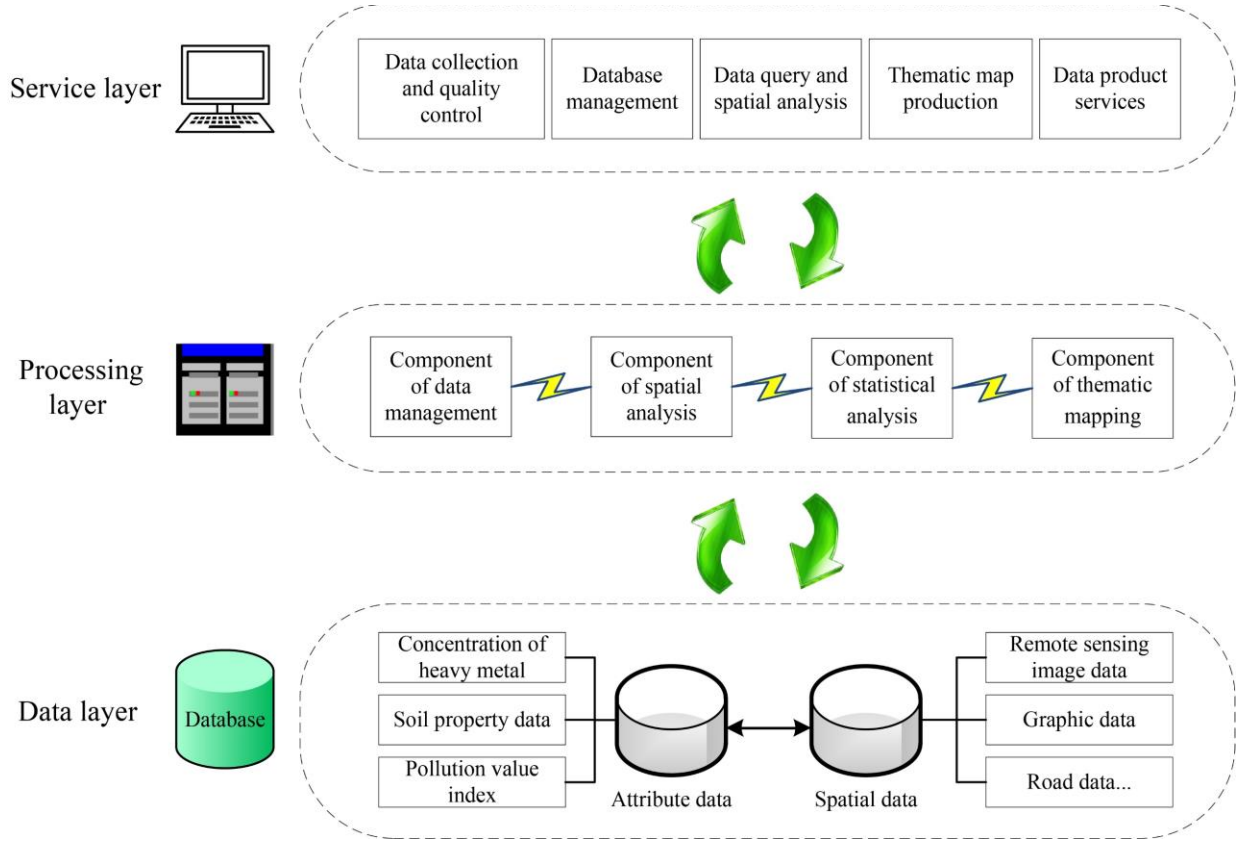


Figure 7. System frame diagram

3）System function design

Map visualization

Generate the point layer according to the coordinates of the sampling points and load it into the current workspace, while loading the required property data in the sheet of the layer. In addition to being used to display map, the map workspace also includes scale, layer management bar, status bar displaying dynamic coordinates and etc. Considering the user's operating habits, the system has the functions of roaming, zooming in, zooming out, selecting and view switching.

Data entry and update

After the database design is completed, the data will be entered into the database through the program interaction interface. The new data entered from the interface will be stored in the attribute database regularly, and the corresponding data will also be entered into the spatial database synchronously. Multimedia data is the data that record the scene, and the system can manage the data that the multimedia tool such as mobile phone, camera obtains. When the original data in the database needs to be updated, the user can modify the data through the database management interface.

Data query and spatial analysis

The query mode of database is divided into conditional and spatial query. Conditional query includes query by unique identifier, attribute of sampling point and value of all cells in the attribute table. Spatial query includes four methods: point, line, rectangle and circle query. The

idea of the algorithm is that after drawing points, lines, circles and rectangles in the workspace, the data content of sampling points within the drawn graph is selected, and the query results are highlighted in the workspace. The results can be exported as a table for editing and sharing.

Soil heavy metals are often collected in area disturbed by human activities, and the samples collected are generally topsoil in the project region. As the location of soil samples is discretized, to obtain the spatial distribution of heavy metals in topsoil, a variable value should be estimated at the unsampled points by using the spatial interpolation method. In order to accurately obtain the spatial distribution of heavy metals in the local high background value area, IDW (Ma et al., 2018) method was selected as the spatial interpolation method embedded in the system.

Data output and thematic map production

Users can use the system to edit the symbol style of the layer, and make thematic maps such as bar chart, pie chart and heavy metal concentration distribution map according to specific requirements. After completing the graphic finishing, the map can be exported to a variety of common picture formats for reference.

4 Discussions

4.1 The main sources of excessive heavy metal concentration

The study showed that heavy metal elements Cd, Cr, Cu and Ni in the project region had at least one sampling point exceeding the standard value respectively. Specially, the exceeding standard rate of total Cd concentration is 100%. Based on PMF analysis, correlation analysis and distribution of soil heavy metal concentration and distribution of pollution source factories, residential area, river and road in the project region, the study identified five sources of soil heavy metals in the project region, including living, traffic, industrial, natural and the agricultural source. As the concentration of Zn and Ag in the project region was lower than the lower limit value of instrumental detection, effective data could not be measured. These two elements were not included in the analysis for the time being, leading to the inability to further identify agricultural source. It was the pity and deficiency of this experiment. Fortunately, there were abundant researches on heavy metal pollution in Pan-Xi region. Combined with their research data and our experimental results, we successfully identified the pollution source.

4.2 The necessity and benefits of building a database platform

Apportionment of the source is only a branch of the research on prevention and control of contamination of soil heavy metals, of which the rest includes the monitoring, evaluation and remediation of soil heavy metal pollution. In order to facilitate the complete transmission of data among different researches of heavy metal pollution and realize the sharing and manageability of data, the establishment of a database platform with powerful data management and sharing capabilities is an option. The database platform of soil heavy metals has the advantages of easy management and sharing. Through the data management interface, users can easily input, update and delete data, and use the modules of query and export quickly to locate requirements and export data, to realize data sharing. When the data stored in the database reach a certain amount, the research that needs long-term data can be supported by the database platform, so that the research can be carried out. For example, remediation of heavy metal pollution and trend analysis dealing with data in different time phases both require support for heterogeneous data. Similarly,

comprehensive data are required for the accurate identification of soil heavy metal pollution source, which needs long-term investigation and monitoring to obtain reliable data.

The database platform performed quite well in providing convenience for management and sharing of data. Data related to soil heavy metal pollution and collected in different regions and at different stages can be stored in the database in an organized way so that researchers can query and export the data through attribute or spatial query. By using the visual interface of the database platform, the spatial distribution of data can be intuitively obtained, as shown in Fig.6. To sum up, it is necessary to establish soil heavy metal database platform for the research of soil heavy metal pollution tracking and remediation.

4.3 Prospect

In the future, in order to give full play to the advantages of data management, the system should have the function for statistical analysis, by which the corresponding analytical results of heavy metal pollution be able to be gotten directly using some built-in analysis method in the system, such as principal component analysis and PMF analysis. Furthermore, based on the stored multi-temporal data, the system can analyze the change trend of heavy metal pollution in the project region over time and provide suggestions for its remediation. The next version of the platform should integrate the management and statistical analysis of heavy metal pollution data, and be able to provide quick analysis results based on pollution tracking and remediation module.

5 Conclusions

A sustainable research process, SSAPD (sustainable source analysis process based on database), for source apportionment of soil heavy metals was explored, including field sampling, indoor data measurement, PMF pollution source identification and analysis, and the construction of soil heavy metal database. The performance of the research process was evaluated by applying it to source identification of soil heavy metals in a mango growing region located at Miyi county, Panzhihua city, and the final conclusions can be summarized as follows:

- Under the guidance of SSAPD, we successfully identified five sources of soil heavy metal pollution in project region, including the living, traffic, industrial, natural and agricultural source.
- A database platform was built, which can realize the management and sharing of data in source identification research of soil heavy metals. Combined with the visual interface of the database platform, it is able to provide convenience for researchers to query and analyze the data stored in database.
- Through management of data by the database platform, the sustainability of the research and data is maintained to a certain extent, enabling researchers to use the data for trend analysis and remediation and other research fields when the amount of data is accumulated enough for the analysis at next stage.

Generally, the SSAPD presented in this article can provide some insights on source apportionment for soil heavy metals. However, in terms of limited availability of stored soil heavy metal data, the advantages of database managing large amounts of multi-temporal data cannot be well demonstrated. Furthermore, the application of this process to the analysis of multi-temporal data is required to improve the reliability.

Acknowledgments, Samples, and Data

This work was supported by National Natural Science Foundation of China (41402248)，the Key R & D projects of Sichuan Science and Technology Department (2018SZ0298), Technology planning projects of the Panzhihua Science and Technology Bureau (2017CY-N-8), the Longshan academic research talent support program of Southwest university of science and technology (17LZX308, 17LZX613, 18LZX638, 18LZXT03), National key research and development projects (2019YFC1803500, 2019YFC1803503, 2019YFC1803504) and the Scientific Research Project of Sichuan Education Department (16ZB0150). The authors also want to thank all the participants of the project for fruitful discussions. The numerical data necessary to replicate the reported results can be found here: https://doi.org/10.5281/zenodo.3841641

References

Barsova, B., Yakimenko, O., Tolpeshta, I., & Motuzova, G. (2019), Current state and dynamics of heavy metal soil pollution in Russian Federation—A review. Environmental Pollution, 249,200-207. doi:10.1016/j.envpol.2019.03.020

Chen, J., Fang, H., Wu, J., Lin, J., Lan, W., & Chen, J. (2019), Distribution and source apportionment of heavy metals in farmland soils using PMF and lead isotopic composition. Journal of Agro-Environment Science, 38(5), 1026-1035. (in Chinese) doi:CNKI:SUN:NHBH.0.2019-05-009

Chen, L., Ni, W., Li, X., & Sun, J. (2009), Investigation of Heavy Metal Concentrations in Commercial Fertilizers Commonly-Used. Journal of Zhejiang Sci-Tech University, 26(2),223-227. (in Chinese) doi：CNKI:SUN:ZJSG.0.2009-02-016

Doabi, S.A., Karami, M., Afyuni, M. and Yeganeh, M. (2018), Pollution and health risk assessment of heavy metals in agricultural soil, atmospheric dust and major food crops in Kermanshah province, Iran. *Ecotox. Environ. Safe.* 163, 153-164. doi：10.1016/j.ecoenv.2018.07.057

Dong, L., Hu, W., Huang, B., Liu, G., Qu, M., & Kuang, R. (2015), Source appointment of heavy metals in suburban farmland soils based on positive matrix factorization. China Environmental Science, 35(7), 2103-2111. (in Chinese) doi：CNKI:SUN:ZGHJ.0.2015-07-031

Eberlv, S. (2005), EPA PMF 1.1 user′s guide[Z]. Washington:U.S. Environmental Protection Agency National Exposure Research Laboratory.

Guan, Q., Wang, F., Xu, C., Pan, N., Lin, J., Zhao, R., Yang, Y., & Luo, H. (2018), Source apportionment of heavy metals in agricultural soil based on PMF: A case study in Hexi Corridor, northwest China. Chemosphere, 193,189-197. doi：10.1016/j.chemosphere.2017.10.151

GB15618-2018. Soil Enwironmental Quality-Risk Control Standard for Soil Contamination of Agricultural Land[S]. *Ministry of Ecology and Environment of the People's Republic of China*. (in Chinese) http://www.mee.gov.cn/ywgz/fgbz/bz/bzwb/trhj/201807/W020190626595212456114.pdf

Gmochowska, W., Pietranik, A., Tyszka, R., Ettler, V., Mihaljevič, M., Długosz, M., & Walenczak, K. (2019), Sources of pollution and distribution of Pb, Cd and Hg in Wrocław soils:

Insight from chemical and Pb isotope composition. Geochemistry, 79(3),434-445. doi:10.1016/j.chemer.2019.07.002

He, J., Yang Y., Christakos, G., Liu, Y., & Yang, X. (2018), Assessment of soil heavy metal pollution using stochastic site indicators. Geoderma, 337,359-367. doi:10.1016/j.geoderma.2018.09.038

Hu, W., Wang, H., Dong, L., Huang, B., Borggaard, O.K., Bruun Hansen, H.C., He, Y. & Holm, P.E. (2018), Source identification of heavy metals in peri-urban agricultural soils of southeast China: An integrated approach. Environ Pollut, 237, 650-661. doi:10.1016/j.envpol.2018.02.070

Huang, Y., Wang, L., Wang, W., Li, T., He, Z., & Yang, X. (2018), Current status of agricultural soil pollution by heavy metals in China: A meta-analysis. Science of The Total Environment, 651(2),3034-3042. doi:10.1016/j.scitotenv.2018.10.185

Han P., Wang Q., & Wang P. (2008), Geographic Information System Development—ArcEngine ways. *Wuhan, Wuhan Unniversity Press*.

Jia, L., Chen, X., & Lv, M. (2013), Heavy metal contamination of soils and its evaluation in different greenbelt regions of Zhangzhou City . Urban Environment ＆ Urban Ecology, (03),11-15. (in Chinese) doi:CNKI:SUN:CHCS.0.2013-03-004

Jorfi, S., Maleki, R., Jaafarzadeh, N., & Ahmadi, M. (2017), Pollution load index for heavy metals in Mian-Ab plain soil, Khuzestan, Iran. Data in Brief, 15,584-590. doi:10.1016/j.dib.2017.10.017

Jiao, W., Chen, W., Chang, A.C., & Page, A.L. (2012), Environmental risks of trace elements associated with long-term phosphate fertilizers applications: a review. Environmental Pollution, 168, 44-53. doi:10.1016/j.envpol.2012.03.052

Khalid, S., Shahid, M., Khan Niazi, N., Murtaza, B., Bibi, I., & Dumat, C. (2017), A comparison of technologies for remediation of heavy metal contaminated soils, Journal of Geochemical Exploration, 182(B),247-268. doi:10.1016/j.gexplo.2016.11.021

Lu, R., & Shi, Z. (1992), Cadmium contents of rock phosphates and phosphate fertilizers of china and their effcets on ecological environment. ACTA PEDOLOGICA SINICA, 29(2), 150-157. (in Chinese) doi:10.1007/BF02677083

Mehr, M.R., Keshavarzi, B., Moore, F., Sharifi, R., Lahijanzadeh, A. & Kermani, M. (2017), Distribution, source identification and health risk assessment of soil heavy metals in urban areas of Isfahan province, *Iran. Journal of African Earth Sciences*, 132, 16-26. doi:10.1016/j.jafrearsci.2017.04.026

Ma, H., Yu, T., Yang, Z., Hou, Q., Zeng, Q., & Wang, R. (2018), Spatial Interpolation Methods and Pollution Assessment of Heavy Metals of Soil in Typical Areas. ENVIRONMENTAL SCIENCE, 39(10), 294-303. doi:10.13227/j.hjkx.201712185

Niu, L.L., Yang, F.X., Chao, X., Yang, H.Y., & Liu, W.P., (2013), Status of metal accumulation in farmland soils across China: from distribution to risk assessment. *Environ. Pollut.* 176, 55-62. doi: 10.1016/j.envpol.2013.01.019

559 Paatero, P., & Tapper, U. (1993), Analysis of different modes of factor analysis as least squares
560 fit problems. Chemom. Intell. Lab. Syst., 18(2),183-194. doi:10.1016/0169-7439(93)80055-M

561 Paatero, P., &Tapper, U. (1994), Positive matrix factorization: a non-negative factor model with
562 optimal utilization of error estimates of data values. Environmetrics, 5(2),111-126.
563 doi:10.1002/env.3170050203

564 Paatero, P. (1997), Least squares formulation of robust non-negative factor analysis. Chemom.
565 Intell. Lab. Syst., 37(1),23-35. doi:10.1016/S0169-7439(96)00044-5

566 Uprety, D., Hejcman, M., Szakova, J., Kunzova, E. & Tlustos, P. (2009), Concentration of trace
567 elements in arable soil after long-term application of organic and inorganic fertilizers. Nutrient
568 Cycling in Agroecosystems, 85, 241-252. doi:10.1007/s10705-009-9263-x

569 Vareda, J.P., Valente, A.J.M., & Duraes, L. (2019), Assessment of heavy metal pollution from
570 anthropogenic activities and remediation strategies. A review. J Environ Manage, 246, 101-118.
571 doi:10.1016/j.jenvman.2019.05.126

572 Wu, J., & Wang. Y. (2000), Progress in Studies on Nutrition an d Fertilization of Non-Pollution
573 Vegetables. Chinese Bulletin of Botany, (06),13-24. (in Chinese) doi: 10.3969/j.issn.1674-
574 3466.2000.06.002

575 Wang, M. (2014). Effects of Iong-term Fertilization on Heavy Metal Accumulationin Soils and
576 Crops[D], Chinese Academy of Agricultural Sciences Dissertation.

577 Wang, P., Li, Z., Liu, J., Bi, X., Ning, Y., Yang, S., & Yang, X. (2019), Apportionment of
578 sources of heavy metals to agricultural soils using isotope fingerprints and multivariate statistical
579 analyses. Environmental Pollution, 249,208-216. doi:10.1016/j.envpol.2019.03.034

580 Wei, Y., Li, G., & Wang, Y, et al. (2018), Investigating factors influencing the PMF model: A
581 case study of source apportionment of heavy metals in farmland soils near a lead-zinc ore.
582 Journal of Agro-Environment Science, 37(11),2549-2559. (in Chinese) doi:10.11654/jaes.2018-
583 0492

584 Xu, J., Chen, Y., & Zhu, Y. (2000), Environmental Analysis of Sustainable Development in Pan-
585 Xi Region ——Using the Accumulation of Cadmium in Soil as an Object. WORLD SCI-TECH
586 R&D, 22(3),39-43. (in Chinese) doi: CNKI:SUN:SJKF.0.2000-03-012

587 Yang, Q., Li, Z., Lu, X., Duan, Q., Huang, L., & Bi, J. (2018), A review of soil heavy metal
588 pollution from industrial and agricultural regions in China: Pollution and risk assessment.
589 Science of The Total Environment, 642,690-700. doi:10.1016/j.scitotenv.2018.06.068

590 Yang, Y., Yang, X., He, M.J., & Christakos, G. (2020), Beyond mere pollution source
591 identification: Determination of land covers emitting soil heavy metals by combining
592 PCA/APCS, GeoDetector and GIS analysis. Catena, 185, 9. doi:10.1016/j.catena.2019.104297

593 Zhou, Y., Yang, D., & Lei, S. (2008), Heavy Metal Content of Soil in Panxi Vegetable
594 Plantations and Their Assessment. SICHUAN ENVIRONMENT, 27(2),67-70 (in Chinese).
595 doi:10.14034/j.cnki.schj.2008.02.016

596 Zhang, Y., Wang, M., Huang, B., Akhtar, M.S., Hu, W., & Xie, E. (2018), Soil mercury
597 accumulation, spatial distribution and its source identification in an industrial area of the Yangtze
598 Delta, China. Ecotoxicol Environ Saf, 163, 230-237. doi:10.1016/j.ecoenv.2018.07.055