Water Supply

IWA PUBLISHING

Water Supply Vol 22 No 4, 3908 doi: 10.2166/ws.2022.020

Investigating the source apportionment of heavy metals in soil surrounding reservoir using partial least-squares regression model

Xu-dong Huang[a,b], Pei-pei Han[c,*], Mei-jing Ma[d], Qiong Cao[e], Wei-zhuo Li[a], Fang Wan[a], Xiao-li Zhang[a], Qi-hui Chai[a], Ling Zhong[a] and Bao-jian Li[a]

[a] College of Water Resources, North China University of Water Resources and Electric Power, Zhengzhou 450046, China
[b] Henan Key Laboratory of Water Resources Conservation and Intensive Utilization in the Yellow River Basin, Zhengzhou 450046, China
[c] Henan Yellow River Hydrological Survey and Design Institute, Zhengzhou 450002, China
[d] Wuhan Institute of Water Science, Wuhan, Hubei 430014, China
[e] No. 2 Institute of Geo-environment Survey of Henan, Zhengzhou 450000, China
*Corresponding author. E-mail: hanpeipei202110@163.com

ABSTRACT

The assessment of heavy metal pollution is crucial for water conservation. This study determined the contents of heavy metals (Cd, Cr, Cu, Ni, Pb, Zn and As) from 39 soil samples surrounding a reservoir, and analyzed the corresponding source and enrichment using enrichment factors and partial least-squares regression. The concentration of Cr (54.06 mg/kg) was lower than the background value of the reservoir area, while the Cd concentration was higher (0.96 mg/kg). Moreover, Cd, Cr, Ni, Pb, Zn and As concentrations in the south exceeded those of the northeast in the Nanwan lake reservoir (NLR). Cd and As were the dominant contaminated elements in the NLR. The Cd enrichment factor value was 11.25, areas with moderate and higher levels of pollution of Cd occupied 89.0% of the total area, while As occupied 18.4%. The dominant sources of Zn, Ni, Cu, Pb and Cr were identified as natural inputs, those of As were agricultural production activities, and those of Cd were industrial production activities. This study provides insight into the heavy metal pollution and key factors of land-use types in watersheds with tea trees as the dominant vegetation cover, and aids in the planning of water pollution prevention and ecological protection.

Key words: enrichment factor method, heavy metals, partial least-squares regression, source apportionment

HIGHLIGHTS

- A novel technology partial least-squares regression model.
- The principle sources were identified by calculating the variable importance for the projection (VIP).
- The inherent defects of traditional regression algorithms in handling multicollinear and noisy data were overcome.

1. INTRODUCTION

Urbanization and industrialization exert significant pressure on the physical and biogeochemical processes in soil and aquatic ecosystems (Resongles *et al.* 2014; Shehab *et al.* 2021). Many studies have been undertaken to investigate the relationships between land use characteristics and the degradation of aquatic compartments by a range of pollutants, including nutrients, organic compounds, and heavy metals (Suresh *et al.* 2012; Wahsha *et al.* 2014; Aiman *et al.* 2016). Heavy metals deserve particular attention as these pollutants are abundant in residential areas, mining operations, factory discharge, sewage systems, etc. and persist in the environment without undergoing biodegradation. Furthermore, heavy metals can be highly toxic, especially in countries where waste management is inadequate in the world (Resongles *et al.* 2014; Ota *et al.* 2020). The enrichment of heavy metals in soil inhibits soil absorption and the corresponding metabolisms, reduces soil nutrient supply, and also affects the yield and quality of agricultural products, directly harming human health through the food chain (Ota *et al.* 2020). Heavy metals and other soil pollutants migrate and enter the water system through the soil-water interface, affecting water quality safety (Yang *et al.* 2021). The degradation of the water environment quality caused by the accumulation and migration of heavy metals in soil is a hot topic in the field of environmental science (Ota *et al.* 2020; Yang *et al.* 2021).

The natural concentrations of heavy metals and metalloids including Cr, Ni, Cd, Cu, Pb and As in soils generally tend to remain low, ensuring an optimum ecological equilibrium. However, the concentrations frequently increase due to human activities, thus adversely affecting many parts of the world (Esmaeili *et al.* 2014; Aiman *et al.* 2016). Land use (e.g. gardens, farms, urban areas, industrial areas, and bare land), which are the most visible form of human activity, plays a pivotal role in the generation, accumulation, and distribution of pollutants amongst air, soil, water, and sediments (Esmaeili *et al.* 2014; Paul *et al.* 2021). Subhani *et al.* (2015) determined As to be the most abundant heavy metal polluter at the surface soil of the Punjab in Pakistan, caused by the rapid development of industrialization and urbanization. Niemitz *et al.* (2013) suggested that sources of excess Cu, and Pb over the background to be fertilizer, pesticides, and other soil amendments. Significant multiple human activities and natural sources of heavy metals present great challenges to the prevention and control of heavy metal pollution (Nicholson *et al.* 2003; Resongles *et al.* 2014). Therefore, quantifying the contributions of changes in individual land use types to different heavy metal composition and enrichment characteristics is of practical importance for providing insights into the heavy metal pollution and its dominant factors, and is essential for water quality security management and planning.

The majority of research employs an integrated approach involving enrichment factors, clustering analysis, geological accumulation indexing, and multivariate regression to relate land use to heavy metals (Suresh *et al.* 2012; Aiman *et al.* 2016; Islam *et al.* 2019). However, particular challenges associated with the use of conventional regression algorithms, namely land use types, are highly co-dependent and are not independent variables (Krishnan *et al.* 2011). Traditional statistical methods have great limitations in solving these problems, making it difficult to reveal the underlying relationship between land use and the spatial and temporal distribution of heavy metals. Therefore, to overcome the inherent defects of traditional regression algorithms in handling multicollinear and noisy data, an extension to the multivariate data analysis technique must be applied (Krishnan *et al.* 2011). Partial least-squares regression (PLSR) methods organically combine data cognition statistical approaches (principal component analysis, canonical correlation analysis) with model-based statistical methods (linear regression analysis) (Krishnan *et al.* 2011; Boongaling *et al.* 2018). PLSR integrates regression modeling, data structure simplification and correlation analysis between two groups of variables under a single algorithm, and is considered as the second-generation regression analysis method (Krishnan *et al.* 2011). As a novel data analysis technique, PLSR has been widely used to overcome multi-collinear and noisy data in many fields for quantitative analyses (Krishnan *et al.* 2011; Boongaling *et al.* 2018; Huang *et al.* 2018; Ndehedehe & Ferreira 2020).

In recent decades, there has been significant concern regarding soil contamination by various toxic metals due to expanding agriculture, industrialization and urbanization surrounding reservoirs (Ye *et al.* 2013). However, heavy metals and metalloids may originate from various sources. Therefore, we must investigate how changes in each land use type influence heavy metals to achieve a more effective and more accurate means of conducting a watershed management approach and to predict heavy metal pollution consequences following land use changes. Nanwan lake reservoir (NLR) is an important potable water source of the Huai river basin, one of the major river systems in China. In recent years, human activities relating to farm, urban, and industrial land around Nanwan lake have increased as a result of heavy metal pollution. In the current study, we integrate field investigations, laboratory analysis, enrichment factor analysis and PLSR to understand the composition and enrichment characteristics of heavy metals and to quantify the individual land use type contributions to major heavy metals. We develop a powerful research strategy to quantify the dominant influencing factors of heavy metals that can act as a basis for environmental planning and decision making at the global scale. The novelty of this study lies in the investigation of the relationships between the partitioning of heavy metals and land use types by employing an integrated approach involving enrichment factors and PLSR. The specific aims are to: (i) provide a scientific basis for research on the potential influence of land use on reservoir water quality; and (ii) provide guidance for the aquatic ecosystem restoration and management of reservoirs.

2. MATERIALS AND METHODS

2.1. Study area

The Nanwan lake reservoir (NLR) is the most important drinking water sources in south of Henan Province (Figure 1). The reservoir area has a typical north subtropical monsoon climate, with an annual average temperature of 15–16 °C and annual average rainfall of approximately 1000 mm. More than 80% of the rainfall mainly occurs from June to October in the monsoon season. The main soil types in the reservoir area are yellow brown soil, including lime soil, paddy soil, and purple soil. In

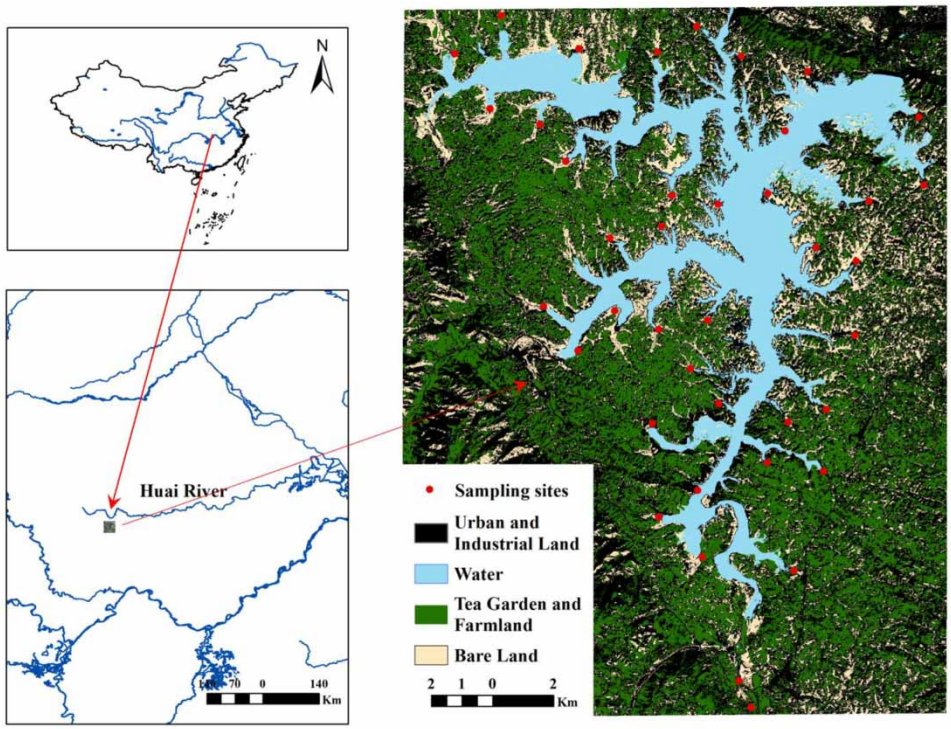

Figure 1 | Location, land use types, and sampling points of the study area.

order to develop the local economy, a large number of economic tree species such as tea were planted around the NLR. Economic development has promoted the expansion of local industrial and urban land. Tea gardens, under both urban and industrial land, may be heavy metals pollution sources for the NLR (Figure 1). Therefore, the NLR is suitable for developing the relationship between heavy metal distribution characteristics and land use.

2.2. Sample collection

Field surveys were divided into the following two stages: (i) the first stage was from 16th May, 2020 to 7th June, 2020, when the Nanwan lake reservoir's water level was 95.75 m; and (ii) the second stage was from 27th June to 17th July, when the reservoir's water level was 96.73–102.40 m. We selected 39 sampling sites (upstream to downstream) based on the geographical characteristics in the NLR. At each site, a sampling plot (1×1 km) was randomly arranged at the elevation from 104 to 120 m (first stage) and 124 to 141 m (second stage). Land use types were investigated and soil samples were collected based on the principle of typical representation and uniform distribution. Following the Technical Specification for Monitoring Farmland Soil Environmental Quality (NY/T395-2000), a handheld GPS was used to determine the location of the sampling points, and an 'S' shaped sampling route was employed to collect 41 surface mixed soil sampling sites with a depth of 0–20 cm in each investigation unit. Amongst these, 13 sampling sites were located within farmland, 14 within urban areas, seven in industrial land, and seven in bare land (Figure 1). All soil samples were taken back to the natural-draft laboratory at room temperature for the removal of impurities (e.g. gravel solids and plant residues). After passing through a 100-mesh nylon sieve, the samples were used for soil heavy metal analysis.

2.3. Test analysis

Soil heavy metal content was measured according to the National Soil Environmental Quality Standard of China (GB 15618-1995). The soil samples were disboiled via the hF-HNO_3 microwave method. Cd, Cr, Cu, Ni, Pb, Zn and Mn concentrations were determined via the ViSTA-MPX plasma emission spectrometer (Varian Inc., UK). As concentrations were determined using the AFS-8220 atomic fluorescence spectrophotometer (Jitian Instrument Company, China). All samples were blank treated, and three repetitions were designed between each sample. In order to ensure the accuracy of the sample analysis,

the national standard soil sample (GSS-1) was used for calibration, and the measurement results were within the allowable error range (<5%).

2.4. Assessment of heavy metal pollution

The enrichment factor method can not only judge the degree of heavy metal pollution according to the enrichment coefficient, but can also quantitatively determine the contribution of anthropogenic and natural sources (Islam *et al.* 2019). The appropriate reference elements must be selected in order to standardize the elements in the sample, thus eliminating the influence of the differences in environmental media, sampling, and sample preparation on the element concentration. Reference elements are rich in content, stable in chemical properties, not affected by human activities and include Ca, Mn, Al and Sc (Loska *et al.* 2004). The Mn content in the study area is reported as 556.90 mg/kg, and its geochemical properties are stable and lower than the background value of soil in the reservoir area. Therefore, Mn is selected as the reference element. The enrichment factor is calculated as follows (Loska *et al.* 2004):

$$EF = [(C_x/C_{Mn})_{sample}]/[(B_x/B_{Mn})_{background}] \quad (1)$$

where EF is the enrichment coefficient of the soil heavy metals; C_x is the measured concentration of the heavy metal elements in the soil; C_{Mn} is the measured concentration of the reference elements in the soil; B_x is the concentration of the heavy metals in the background environment; and B_{Mn} is the concentration of the reference elements in the background environment. According to Sutherland (2000), heavy metal pollution can be divided into five levels via the enrichment coefficient (Table 1).

We employed the soil environmental background value of Henan Province as the soil background value within the reservoir area, with Cd, Cr, Cu, Ni, Pb, Zn, As and Mn contents of 0.07, 62.50, 19.20, 26.10, 19.10, 58.40, 10.90 and 567.00 mg/kg, respectively, according to the China National Environmental Monitoring Center (1990). In order to ensure the authenticity of the evaluation results, the soil environmental background value was selected as the reference standard for the enrichment degree of heavy metals in the reservoir area.

2.5. Statistical analysis

Prior to the statistical analysis of the data, we tested the normality of the data distribution using the Kolmogorov-Smirnov test (KST) at the 5% significance level. Such tests have been widely used to verify the fitting of probability distributions in the field of water science and typically prioritize the behavior of the probability in the distribution tails (Beskow *et al.* 2015). The KST is based on the largest difference, in absolute value, between the theoretical and empirical cumulative probabilities. The maximum deviation is characterized as follows:

$$D_{\max} = \text{maximum } |F(X) - Fn(X)| \quad (2)$$

The bilateral probability associated with the occurrence (under H_0) of values as great as the value of the observed D, will be determined according to the table of critical values. Therefore, if $|D_{\max}| \geq D_{\text{critical, a,}}$ the null hypothesis is rejected, which means that the sample has a different distribution to that of the tested theoretical distribution (Beskow *et al.* 2015). The test results are reported in Table 2.

Table 1 | Enrichment factor grades

Enrichment coefficient (EF)	Pollution levels	The degree of pollution
EF < 2	1	EF < 1 is pollution-free;1 < EF < 2 is mild pollution
2 < EF < 5	2	Moderate pollution
5 < EF < 20	3	High levels of pollution
20 < EF < 40	4	Serious pollution
EF < 40	5	Severe pollution

Table 2 | Descriptive statistical summary of heavy metal content in soils ($n = 41$)

Heavy metal	Ranges (mg/kg)	Average	Standard deviation	Coefficient of variation (CV)	Partial degrees	kurtosis	K-s value	Distribution type	Soil environmental background value (mg/kg)
Cd	0.05–7.25	0.96	0.78	0.58	1.07	1.69	1.24	normal	0.07
Cr	4.23–158.56	54.06	42.96	0.76	0.20	0.65	0.86	normal	62.50
Cu	7.12–84.54	23.22	12.45	0.55	0.28	0.57	0.95	normal	19.20
Ni	3.85–194.58	31.98	17.46	0.65	0.22	0.06	0.78	normal	26.10
Pb	4.76–154.38	26.89	19.65	0.71	0.25	1.21	0.84	normal	19.10
Zn	31.87–165.45	70.52	21.89	0.39	0.34	0.45	0.84	normal	58.40
As	0.37–139.64	11.08	12.41	1.34	0.63	0.87	0.74	normal	10.90
Mn	122.84–875.12	557.76	117.49	0.17	0.97	2.35	0.79	normal	567.00

PLSR is a multivariate iterative projection approach used to model the quantitative relationship between different heavy metals and land use areas in our study. Dependent variable matrix F is established for heavy metal concentration Y_j, and the independent variable matrix E is established for land use area X_i. Both X and Y are simultaneously modeled to determine the latent variables in X that best predicts the latent variables in Y. The basic PLSR algorithm combines and generalizes the features from principal component analysis (PCA) and multiple linear regression. Based on PCA, first principal component T of the independent variable matrix is calculated to obtain component vector tE. The first principal component U of the dependent variable matrix is determined, and the correlation between U and T is calculated using the canonical correlation analysis principle. The residual matrix following the extraction of the first principal component can be continued to extract the second component, and canonical correlation analysis is performed until all the correlations are decomposed. The ordinary least squares regression equations of dependent variable matrix $F(Y)$ and score vector tE are established respectively, and the ordinary least squares regression equations of tE and independent variable matrix $E(X)$ are substituted into the ordinary least squares regression equations of tE and independent variable matrix $E(X)$. Finally, PLSR models for heavy metals are constructed to identify the main land use controlling factors. Improving on ordinary PCA, PLSR uses only the most important linear combinations and provides prediction results with a higher accuracy. The regression coefficient (RC) indicates the direction and strength of the impact of each variable in the PLSR model. The contribution of each explanatory variable in fitting the model can be described by the Variable Influence on Projection (VIP), derived as the sum of square of the PLSR weights across all components (Huang *et al.* 2018). For interpretation purposes, variables with higher VIP values are considered as more important (Krishnan *et al.* 2011). More details on PLSR can be found in Krishnan *et al.* (2011).

3. RESULTS AND DISCUSSION

3.1. Distribution characteristics of heavy metal content in soil

The heavy metal contents of soil in the NLR are reported in Table 2 and Figure 2. Compared with the background value of the soil environment in the reservoir area, Cr (Cd) content in the NLR was lower (higher) than the background value. The higher content of Cd may be related to the input of pollution load caused by the discharge of industrial wastewater around the reservoir, as well as the improper disposal of garbage left over by migration and relocation. In addition, reservoir functions (e.g. flood control, irrigation, power generation, breeding, and tourism) are greatly affected by human interference, which can easily cause Cd enrichment. These results are consistent with the study of toxic metals in Pakistan by Aiman *et al.* (2016). The content of Cu, Ni, Pb, Zn, and As in the NLR were similar to those of the background. Ni, Pb, Zn, and As exhibited higher content in the northeast of NLR compared to the south. This can be attributed to the flatter terrain in the northeast of NLR than the south, with a higher population density and faster growing economy. The enrichment of heavy metals by intensive human activity is more obvious. These results are similar with Nicholson *et al.* (2003) and Wahsha *et al.* (2014), who argued that the enrichment of Ni, Pb, Zn, and As are mainly attributed to human activity. The variation coefficient of Cd, Cr, Cu, Ni, Pb, Zn, and As ranged from 0.39 to 1.34, with the latter exhibiting the largest value (Table 2). The minimum, maximum, and average content of As were lower in the northeast compared to the south, and its enrichment was generally concentrated in the south (Figure 2). This may be related to mining and sewage irrigation, as well as pesticide and fertilizer

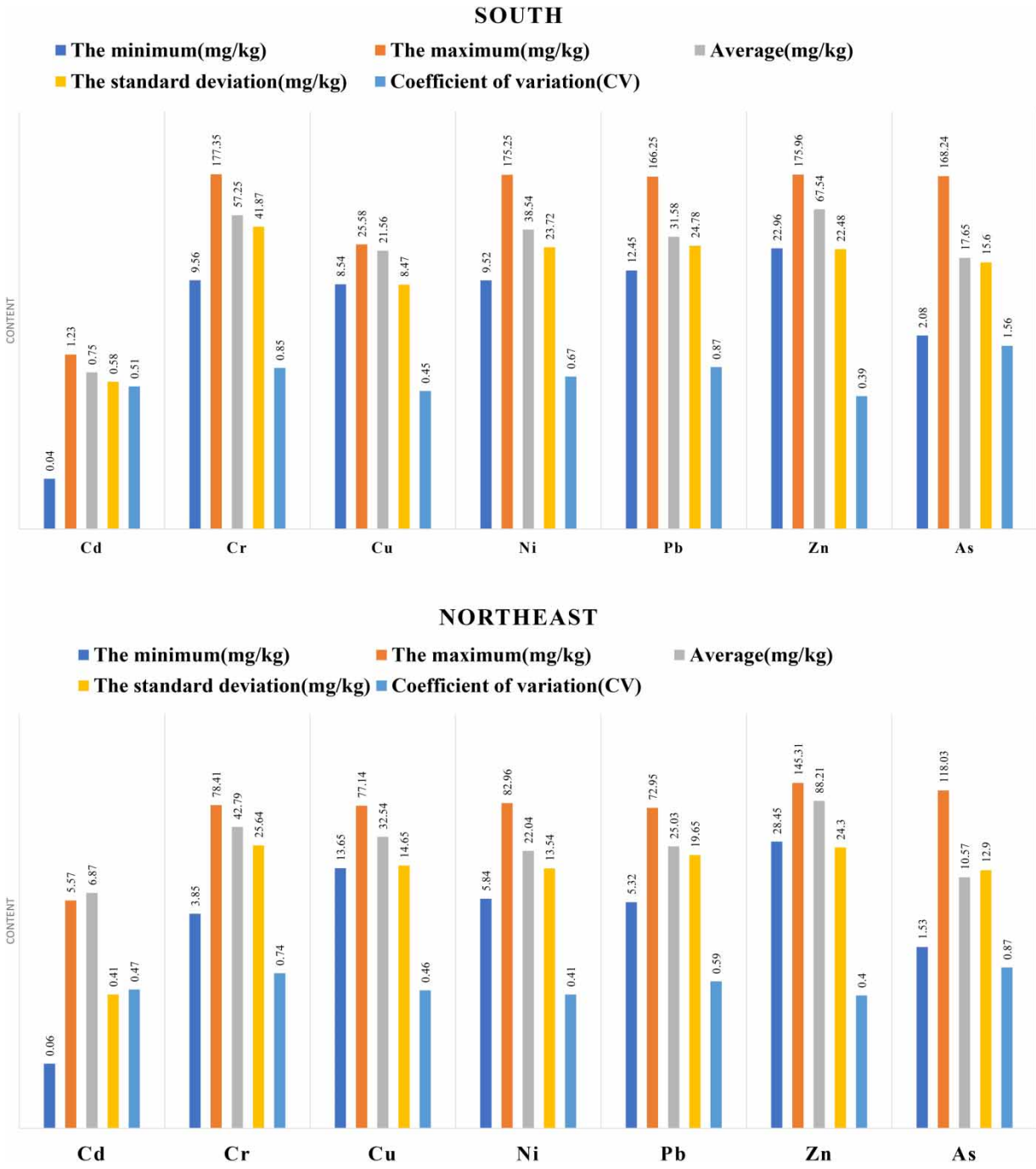

Figure 2 | Heavy metal content in the reservoir ($n = 41$).

applications in agricultural activities. In the south of the NLR area, the cultivation of tea requires the application of a large amount of phosphate fertilizer, which contains arsenic. In addition, the wharf in this region may also enhance the As content in the south of the NLR area. These conclusions are consistent with Resongles *et al.* (2014), who suggested that Cd and As enrichment in soils may be related to the mining industry, smelting industry and agricultural fertilization.

3.2. Enrichment level of heavy metals in soil

The heavy metal enrichment coefficients in the NLR are presented in Table 3. The average enrichment coefficients of Cd, Cr, Cu, Ni, Pb, Zn and As were determined as 11.25, 0.62, 1.45, 1.21, 1.11, 1.52 and 1.40, respectively, most of which are close to

Table 3 | Distribution of heavy metal enrichment coefficients of soil

Heavy metal	Enrichment factor coefficient
Cd	11.25
Cr	0.62
Cu	1.45
Ni	1.21
Pb	1.11
Zn	1.52
As	1.40

1.0. This reveals the strong influence of natural sources. Moreover, the enrichment coefficient of Cd exceeds 10.0, indicating that the Cd enrichment in the study area is not only from natural source substances, but is also influenced by human activities (e.g. domestic pollution and pollutant discharge in industrial production). These results are in agreement with Suresh *et al.* (2012), who proposed the mining industry, agricultural production activities, pollutant discharging in industrial production, and the domestic garbage of surrounding residents to be key influencing factors for the excess of Cd. Ye *et al.* (2013) investigated heavy metals in the fluctuation zone of the Three Gorges reservoir, and indicated that the high risk of Cd was generally caused by the discharge of industrial pollutants. In summary, Cd has become a major ecological risk factor with complex sources related to human activities. Thus, relevant departments should focus more attention on these sources to strengthen their prevention and control.

The evaluation of the enrichment factors reveals that 85% of the study area exhibited no pollution to slight pollution of Cr, Cu, Zn, Pb and Ni. Furthermore, 18.4 and 89.0% of the area exhibited moderate (or greater) As and Cd pollution, respectively (Table 4). Cd enrichment was identified as the most serious pollution element in the NLR, followed by As. A comparison with the ecological assessment results of heavy metals from the same data sources reveals serious Cd pollution across the study area. This is in agreement with Ye *et al.* (2013), who employed the geological accumulation index to evaluate soil heavy metals in the water-level fluctuation zone in the Three Gorges. The authors detected heavy Cd pollution, with agricultural production activities as the dominant source, while As was associated with a low potential risk. Zhang *et al.* (2013) determined As to be the most abundant heavy metal polluter at the surface soil of the Qingshan reservoir, caused by the rapid development of industrialization and urbanization. Previous research demonstrates the great impact of human activity on heavy metals in the surface sediments and the surrounding soil of each reservoir. Therefore, it is necessary to analyze the sources of heavy metals by evaluating their enrichment characteristics.

3.3. Correlation analysis of heavy metals in soil

The correlation analysis results of the heavy metals in the NLR soil indicate a strong correlation between the heavy metals. The correlation coefficients of Cr, Cu, Ni, Pb and Zn ranged from 0.445 to 0.753, reaching a strong significant positive correlation (Table 5, $P < 0.01$). Wahsha *et al.* (2014) demonstrated Cu, Pb and Zn to be strongly correlated, indicating significant homology, and have similar geochemical behavior to the Cr and Ni. The degree of pollution with these metals was affected by

Table 4 | Percentage of heavy metals across different pollution levels (%)

	The percentage						
Pollution level	Cd	Cr	Cu	Ni	Pb	Zn	As
EF < 2	11.0	97.2	87.5	88.8	90.0	86.1	81.6
2 < EF < 5	8.4	2.8	10.2	10.8	10.0	13.9	17.6
5 < EF < 20	69.9	–	2.3	0.4	–	–	0.8
20 < EF < 40	8.9	–	–	–	–	–	–
EF < 40	1.8	–	–	–	–	–	–

Table 5 | Coefficient correlations between heavy metals and soil properties

Heavy metals	Cd	Cr	Cu	Ni	Pb	Zn	As
Cd	1						
Cr	0.120	1					
Cu	0.142	0.487**	1				
Ni	0.055	0.714**	0.678**	1			
Pb	0.064	0.753**	0.454**	0.578**	1		
Zn	0.214*	0.448**	0.485**	0.565**	0.445**	1	
As	0.221*	0.278**	0.347**	0.312**	0.277**	0.124	1
Total phosphorus	0.174	−0.057	0.059	−0.038	−0.123	0.068	0.203*

Note: *significant correlation at 0.05 level; **significant correlation at 0.01 level ($n = 41$).

the intrinsic properties of the soils and human activity. The Cu and Zn in farmland soils is affected by agricultural activities, such as fertilization and wastewater irrigation (Nicholson *et al.* 2003). The correlation coefficient between As and total phosphorus was 0.203, reaching a significant correlation level ($P < 0.05$), while that between Cd and the other six heavy metals was not significant. As and total phosphorus were homologous due to the application of phosphate fertilizer which contained arsenic (Nicholson *et al.* 2003). Wahsha *et al.* (2014) revealed Cd, Cu, Pb and Zn to be strongly homologous, due to its unique geographical location of suburban farmlands, with a dense road network and its role as an important vegetable and grain production area for urban residents. However, our results suggested that the correlation coefficient between Cd and the other six heavy metals was not significant in the NLR. Cd in soil may be derived from vehicle exhaust emissions (Wahsha *et al.* 2014). The tea gardens in the study area were planted in terraced fields across rural-urban fringes, located far away from areas with dense road networks. Thus, Cd and the other six heavy metals exhibited spatial heterogeneity in our study area.

3.4. Factor analysis of soil heavy metals

Kaiser-Meyer-Olkin and Bartlett sphericity tests were used to assess the validity and reliability of the data and are typically performed prior to the factor analysis (Cortes *et al.* 2021). According to Ye *et al.* (2013), the corresponding results of $0.81 > 0.7$ and 352.16 ($DF = 21$, $P < 0.01$) determined in our study indicate a strong linear correlation between the seven heavy metals, suggesting that the factor analysis was effective. Table 6 reports the factor analysis results, where the first three factors reflect 78.5% of the total variance of the seven heavy metals, and can explain the majority of the heavy metal source information. Factor 1 includes Cr, Cu, Zn, Pb and Ni, with a variance contribution rate of 46.8%; factor 2 contains As, with a variance contribution rate of 17.4%; and factor 3 contains Cd, with a contribution rate of 14.4%. These results indicated that Cr, Cu, Zn, Pb and Ni could be considered as similar dependent variables, with a similar proportion of

Table 6 | Varimax rotated factor loading soil heavy metals for farmland soil in the NLR

	Rotate the pre-transformation factor			Rotate the transformation factor		
Heavy metal	1	2	3	1	2	3
Cd	0.235	0.852	0.378	0.015	0.156	0.975
Cr	0.785	0.146	0.097	0.945	0.185	0.056
Cu	0.602	0.076	0.185	0.656	0.421	0.078
Ni	0.875	0.167	0.089	0.845	0.365	0.124
Pb	0.849	0.198	0.028	0.985	0.323	0.015
Zn	0.628	0.025	0.685	0.699	0.345	0.452
As	0.568	0.476	0.752	0.256	0.754	0.258
Accumulated load (%)	50.25	67.25	78.54	46.81	64.16	78.54

explained variability and cumulative explained in land use type, and similar main controlling factor in our subsequent PLSR models.

3.5. PLSR analysis of heavy metals in the soil

Table 7 summarizes the PLSR model results. The first-order components of the Cr, Cu, Ni, Pb and Zn models included bare land, tea garden and farmland on the positive and explained 62.7–71.2% of the variation of heavy metals. The addition of the second-order components (bare, urban and industrial land), augmented the model-explained variance to 71.5–76.2% and generated minimum RMSECV values (1.25–3.65). The addition of more components did not promote the explanation substantially but led to higher RMSECV values, indicating that the subsequent components were not significantly correlated with the residuals of the predicted variable (Carrascal *et al.* 2010) (Table 7). VIP values for bare land were greater than 1, while tea gardens, farmland, urban and industrial land had VIP values smaller than 1 (Figure 3). According to Islam *et al.* (2019), Cr, Cu, Ni, Pb and Zn can be considered as natural-related enrichment metals in soils. Bare land, which was less affected by human activity, was the most important land use types for Cr, Cu, Ni, Pb and Zn. In the As model, the first-order component was tea garden and farmland on the positive, and explained 71.4% of the As variance (Table 7). The addition of the second-order and third-order components (agricultural and industrial land) on the positive augmented the model-explained variance to 86.9% (Table 7). As shown in Figure 3, tea garden and farmland had VIP values larger than 1, while VIP values lower than 1 were observed for urban and industrial land. The results indicate that agricultural activities such as fertilization and pesticide usage lead to arsenic enrichment. In the Cd model, the first-order component was urban and industrial land on the positive, and explained 69.5% of the Cd variance (Table 7). The addition of the second-order and third-order components (tea garden and farmland), augmented the model-explained variance to 83.3% (Table 7). As shown in Figure 3, urban and industrial land

Table 7 | Summary of PLSR models for heavy metals in the reservoir area

Response Y	R^{2a}	Component	% of explained variability in Y	Cumulative explained in *Y* (%)	RMSECV
Cd	0.83	1	69.5	69.5	11.25
		2	10.2	79.7	2.53
		3	3.6	83.3	2.47
		4	0.8	84.1	4.65
Cr	0.75	1	71.2	71.2	3.56
		2	3.7	74.9	2.89
		3	0.9	75.8	3.04
Cu	0.76	1	67.9	67.9	7.58
		2	8.3	76.2	3.02
		3	1.0	77.2	4.87
Ni	0.72	1	65.9	65.9	8.12
		2	5.6	71.5	1.25
		3	0.6	72.1	2.98
Pb	0.72	1	62.7	62.7	8.45
		2	9.4	72.1	3.65
		3	2.5	74.6	4.18
Zn	0.74	1	68.4	68.4	7.64
		2	5.9	74.3	2.69
		3	0.6	74.9	4.78
As	0.87	1	71.4	71.4	9.54
		2	12.0	83.4	4.12
		3	3.5	86.9	3.24
		4	0.4	87.3	4.29

[a]R^2 is goodness-of-fit.

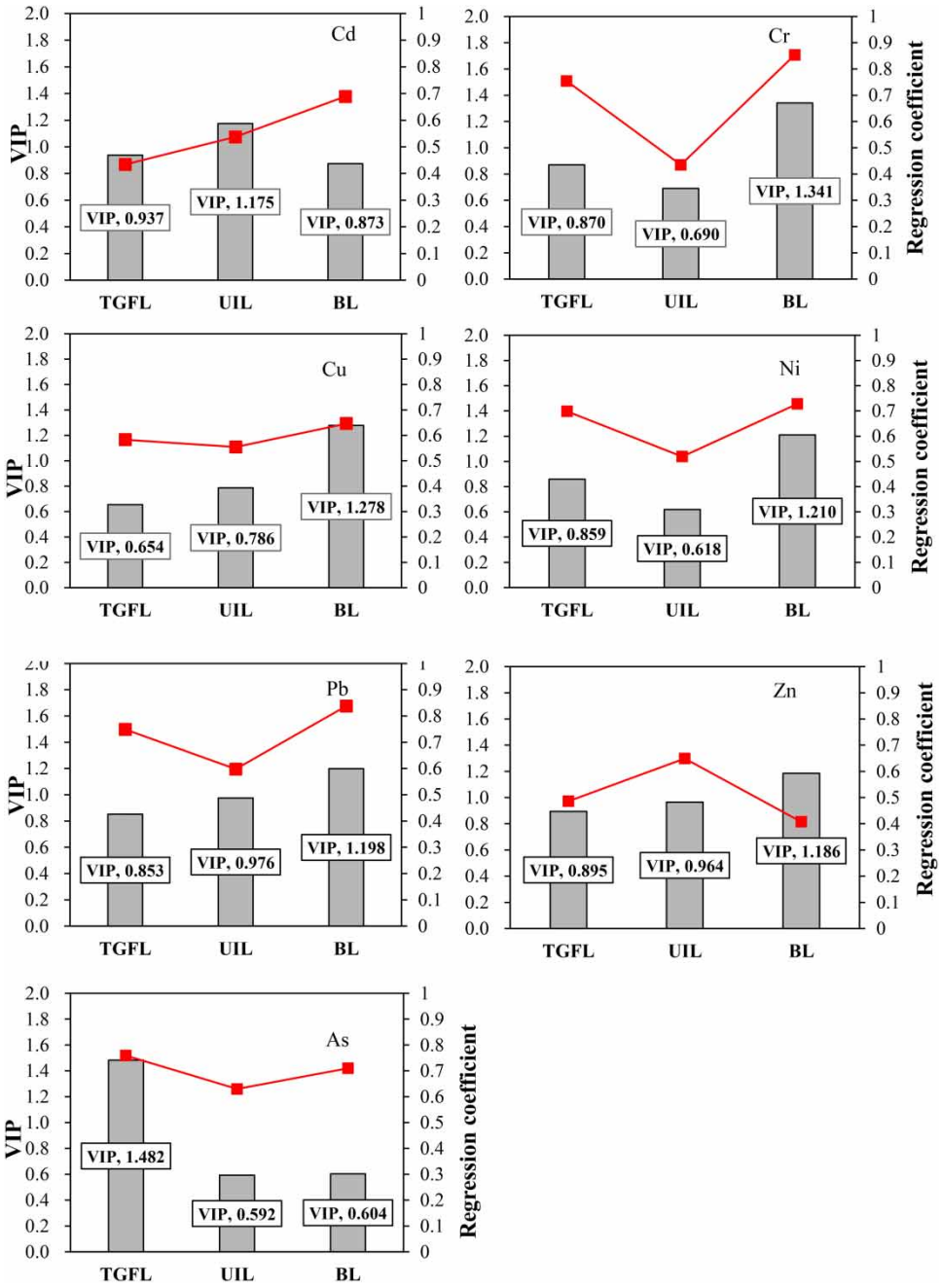

Figure 3 | VIP and regression coefficients for land use types (sampling TGFL = tea garden and farmland; UIL = urban and industrial land; BL = bare land).

exhibited VIP values larger than 1, while VIP values lower than 1 were observed for tea garden and farmland. Amongst the possible sources of Cd (residential areas, industrial areas, mining and agricultural activities), our results suggest that the effects of urban and industrial land were much greater than that of agricultural areas due to large and dense populations, the relocation of waste, factory sites, and the existence of tungsten mines, as well as Cd-containing pollutant discharge from the power station and wharf. Similar results were reported by Esmaeili *et al.* (2014) and Aiman *et al.* (2016).

3.6. Source analysis of soil heavy metals

Based on the PLSR analysis, the sources of the seven target heavy metals in the NLR soil can be divided into three categories:

(1) Natural inputs. Our study revealed the dominant sources of Cr, Cu, Ni, Pb and Zn to be natural inputs, with human activities having less of an impact (Ye *et al.* 2013). In particular, strong relationships were observed between Cr, Cu, Zn, Pb and bare land, which is less affected by human activity. The Cr, Cu, Ni, Pb and Zn content in the NLR were lower those of the background values of the reservoir soil, while there was less of a difference between those of Cu, Ni, Pb and Zn and the reservoir background values, with EF values less than 1.6. Moreover, 85.0% of the samples in the study area were pollution-free to slightly polluted. The correlation analysis of the heavy metals revealed a significant correlation among the five heavy metals ($P < 0.01$). Pb, Cu and Zn are cuprophilic elements, and have similar geochemical behaviors in soil (Zhang *et al.* 2013). In summary, the five heavy metals belong to the same category (category 3), and originate from natural inputs.

(2) Agricultural production activities. Table 7 demonstrates a significant positive correlation between As and agricultural land (tea garden and farmland). This is attributed to the application of phosphorus fertilizer, which contains trace As and is an important source of As in the soil (Signes-Pastor *et al.* 2007). The As content in the NLR is 1.28 mg/kg higher than the soil background value of the reservoir area, and 18.4% of the samples exhibit a moderate (or greater) pollution degree. Moreover, the content of As in the south exceeds that in the northeast due to the larger agricultural land area and the relatively prominent non-point source pollution. Therefore, the enrichment of As is likely to be related to the application of pesticides and chemical fertilizers, unreasonable farming methods and sewage irrigation.

(3) Industrial engineering and domestic waste. The average content of Cd was 14.8 times that of the soil environmental background value. The Cd content in the northeast was higher than that in the south, which can be attributed to the small population and slow economic development in the latter. The majority of the study area contains suburbs or rural land, and thus Cd pollution is most likely related to the large rural area and dense population covered by the Nanwan lake watershed. Following the relocation of a large number of waste, factory sites and other incomplete cleaning processes, the rain erosion and surface runoff is discharged into the nearby soil (Aiman *et al.* 2016). Furthermore, the existence of tungsten mines in Henan Province, as well as the power station and wharf around the reservoir area, are extremely likely to cause Cd enrichment.

The comprehensive application of clustering analysis, geological accumulation indexing and multiple regression methods may seem like a reasonable strategy to generate a mathematical model that relates land use to heavy metals. However, our results showed that the use of conventional regression algorithms to investigate the relative importance of land use is limited as heavy metals and land-use types are highly co-dependent. Therefore, our research employed the coupling of enrichment factors and partial least-squares regression to handle multicollinear and noisy data. This is a novel research strategy to investigate heavy metal pollution and the corresponding dominant land-use types in watersheds, and to provide useful information for water pollution prevention and ecological protection.

4. CONCLUSION

In the current study, we focus on the contents and source of heavy metals surrounding a reservoir. We analyzed the corresponding source and enrichment of Cd, Cr, Cu, Ni, Pb, Zn and As based on the enrichment factor (EF) method, and quantified the relative importance of land use types on these heavy metals based on PLSR models and VIP values. The results indicate the Cr content in the NLR to be lower than that of the background value of the reservoir area, while the Cd content is higher. Moreover, unlike Cu and Cd, the contents of Cr, Ni, Pb, Zn and As in the south exceed those in the northeast. We determine 85% of the NLR to exhibit pollution-free to slightly polluted levels of Cr, Cu, Zn, Pb and Ni; 89.0% of the study area has moderate (or greater) pollution of Cd; 18.4% presents moderate (or greater) As pollution. Thus, the enrichment of Cd and As is a key problem in the study area.

Heavy metals in the farmland soil of the NLR are generally attributed to natural sources, agricultural production activities and industrial and domestic wastes, with contribution rates of 46.8, 17.4 and 14.4%, respectively. Cr, Cu, Ni, Pb and Zn can be considered to have natural sources, with bare land dominating tea gardens, farmland, urban land and industrial land as the key source. VIP values of tea gardens and farmland exceeded those for urban and industrial land for As, indicating that agricultural activities lead to arsenic enrichment. For Cd, VIP values of urban and industrial land are larger than those of tea gardens and farmland, suggesting that the effects of urban and industrial land are much greater than those of agricultural areas. Cd enrichment is mainly affected by dense populations, waste, factories, mines, power stations and the wharf, which discharges pollutants containing Cd.

We developed a novel strategy to determine heavy metal pollution and the dominant land-use types in watersheds, effectively overcoming the spatial collinearity of different land use types. By revealing the underlying relationship between heavy metals and land use factors, our work can provide useful information for water pollution prevention and ecological protection.

ACKNOWLEDGEMENTS

Financial support provided by PhD early development program of North China University of Water Resources and Electric Power.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

REFERENCES

Aiman, U., Mahmood, A., Waheed, S. & Malik, R. N. 2016 Enrichment, geo-accumulation and risk surveillance of toxic metals for different environmental compartments from Mehmood Booti dumping site, Lahore city, Pakistan. *Chemosphere* **144**, 2229. https://doi.org/10.1016/j.chemosphere.2016.02.009.

Beskow, S., Calderia, T. L., De Mello, C. R., Faria, L. C. & Guedes, H. A. S. 2015 Multiparameter probability distributions for heavy rainfall modeling in extreme southern Brazil. *Journal of Hydrology-Regional Studies* **4** (B), 123–133. https://doi.org/10.1016/j.ejrh.2015.06.007.

Boongaling, C., Faustino-Eslava, D. V. & Lansigan, F. P. 2018 Modeling land use change impacts on hydrology and the use of landscape metrics as tools for watershed management: the case of an ungauged catchment in the Philippines. *Land Use Policy* **72**, 116–128. https://doi.org/10.1016/j.landusepol.2017.12.042.

Carrascal, L., Galván, I. & Gordo, O. 2010 Partial least squares regression as an alternative to current regression methods used in ecology. *Oikos* **118** (5), 681–690. https://doi.org/10.1111/j.1600-0706.2008.16881.x.

Cortes, S., Burgos, S., Adaros, H., Lucero, B. & Quiros-Alcala, L. 2021 Environmental health risk perception: adaptation of a population-based questionnaire from Latin America. *International Journal of Environmental Research and Public Health* **18** (16), 8600. https://doi.org/10.3390/ijerph18168600.

Esmaeili, A., Moore, F., Keshavarzi, B., Jaafarzadeh, N. & Kermani, M. 2014 A geochemical survey of heavy metals in agricultural and background soils of the Isfahan industrial zone, Iran. *Catena*. https://doi.org/10.1016/j.catena.2014.05.003.

Huang, X., Wang, L., Han, P. & Wang, W. 2018 Spatial and temporal patterns in nonstationary flood frequency across a forest watershed: linkage with rainfall and land use types. *Forests* **9** (6), 339. https://doi.org/10.3390/f9060339.

Islam, M. S., Ahmed, M. K., Al-Mamun, M. H. & Islam, S. M. A. 2019 Sources and ecological risks of heavy metals in soils under different land uses in Bangladesh. *Pedosphere* **29** (05), 123–133. https://doi.org/10.1016/S1002-0160(17)60394-1.

Krishnan, A., Williams, L. J., McIntosh, A. R. & Abdi, H. 2011 Partial least squares (pls) methods for neuroimaging: a tutorial and review. *NeuroImage* **56** (2), 455–475. https://doi.org/10.1016/j.neuroimage.2010.07.034.

Loska, K., Wiechuła, D. & Korus, I. 2004 Contamination of metals by farming soils affected by industry. *Environment International* **30** (2), 159–165.

Ndehedehe, C. E. & Ferreira, V. G. 2020 Assessing land water storage dynamics over South America. *Journal of Hydrology* **580**, 124339. https://doi.org/10.1016/j.jhydrol.2019.124339.

Nicholson, F. A., Smith, S. R. & Alloway, B. J. 2003 An inventory of heavy metal inputs to agricultural soils in England and Wales. *The Science of the Total Environment* **311**, 205–219. https://doi.org/10.1016/S0048-9697(03)00139-6.

Niemitz, J., Haynes, C. & Lasher, G. 2013 Legacy sediments and historic land use: chemostratigraphic evidence for excess nutrient and heavy metal sources and remobilization. *Geology* **41** (1), 47–50. https://doi.org/10.1130/G33547.1.

Ota, Y., Suzuki, A., Yamaoka, K., Nagao, M., Tanaka, Y., Irizuki, T., Fujiwara, O., Yoshioka, O., Kawagata, S., Kawano, S. & Nishimura, O. 2020 Geochemical distribution of heavy metal elements and potential ecological risk assessment of Matsushima Bay sediments during 2012–2016. *Science of the Total Environment* **751**, 141825. https://doi.org/10.1016/j.scitotenv.2020.141825.

Paul, V., Sankar, M. S., Vattikuti, S., Dash, P. & Arslan, Z. 2021 Pollution assessment and land use land cover influence on trace metal distribution in sediments from five aquatic systems in southern USA. *Chemosphere* **263**, 128243. https://doi.org/10.1016/j.chemosphere.2020.128243.

Resongles, E., Casiot, C., Freydier, R., Dezileau, L., Viers, J. & Elbaz-Poulichet, F. 2014 Persisting impact of historical mining activity to metal (Pb, Zn, Cd, Tl, Hg) and metalloid (As, Sb) enrichment in sediments of the Gardon River, southern France. *Science of the Total Environment* **481**, 509–521. https://doi.org/10.1016/j.scitotenv.2014.02.078.

Shehab, Z. N., Jamil, N. R., Aris, A. Z. & Shafie, N. S. 2021 Spatial variation impact of landscape patterns and land use on water quality across an urbanized watershed in Bentong, Malaysia. *Ecological Indicators* **122**, 107254. https://doi.org/10.1016/j.ecolind.2020.107254.

Signes-Pastor, A., Burló, F., Mitra, K. & Carbonell-Barrachina, A. 2007 Arsenic biogeochemistry as affected by phosphorus fertilizer addition, redox potential and pH in a west Bengal (India) soil. *Geoderma* **137** (3-4), 504–510. https://doi.org/10.1016/j.geoderma.2006.10.012.

Subhani, M., Mustafa, I., Alamdar, A., Katsoyiannis, I. A., Ali, N. & Huang, Q. 2015 Arsenic levels from different land-use settings in Pakistan: bio-accumulation and estimation of potential human health risk via dust exposure. *Ecotoxicology & Environmental Safety* **115**, 187–194. https://doi.org/10.1016/j.ecoenv.2015.02.019.

Suresh, G., Sutharsan, P., Ramasamy, V. & Venkatachalapathy, R. 2012 Assessment of spatial distribution and potential ecological risk of the heavy metals in relation to granulometric contents of Veeranam lake sediments, India. *Ecotoxicology and Environmental Safety* **84**, 117–124. https://doi.org/10.1016/j.ecoenv.2012.06.027.

Sutherland, R. 2000 Bed sediment-associated trace metals in an urban stream, Oahu, Hawaii. *Environmental Geology* **39** (6), 611–627. https://doi.org/10.1007/s002540050473.

Wahsha, M., Bini, C., Zilioli, D., Spiandorello, M. & Gallo, M. 2014 Potentially harmful elements in terraced agroecosystems of NE Italy: geogenic vs anthropogenic enrichment. *Journal of Geochemical Exploration* **144**, 355–362. https://doi.org/10.1016/j.gexplo.2014.01.012.

Yang, S., Feng, W., Wang, S., Chen, L., Zheng, X., Li, X. & Zhou, D. 2021 Farmland heavy metals can migrate to deep soil at a regional scale: a case study on a wastewater-irrigated area in China. *Environmental Pollution* **281**, 116977. https://doi.org/10.1016/j.envpol.2021.116977.

Ye, C., Li, S., Zhang, Y., Tong, X. & Zhang, Q. 2013 Assessing heavy metal pollution in the water level fluctuation zone of China's three Gorges Reservoir using geochemical and soil microbial approaches. *Environmental Monitoring and Assessment* **185** (1), 231–240. https://doi.org/10.1007/s10661-012-2547-7.

Zhang, F., Yang, C. & Pan, R. 2013 Pollution characteristics and ecological risk assessment of heavy metals in surface sediments of Qingshan Reservoir in Lin' an City, Zhejiang Province of East China. *Chinese Journal of Applied Ecology* **24** (9), 2625–2630. (in Chinese).

First received 11 October 2021; accepted in revised form 31 December 2021. Available online 19 January 2022