Chapter

Data Mining for Source Apportionment of Trace Elements in Water and Solid Matrix

Yao Shan and Jianjun Shi

Abstract

Trace elements migrate among different environment bodies with the natural geochemical reactions, and impacted by human industrial, agricultural, and civil activities. High load of trace elements in water, river and lake sediment, soil and air particle lead to potential to health of human being and ecological system. To control the impact on environment, source apportionment is a meaningful, and also a challenging task. Traditional methods to make source apportionment are usually based on geochemical techniques, or univariate analysis techniques. In recently years, the methods of multivariate analysis, and the related concepts data mining, machine learning, big data, are developing fast, which provide a novel route that combing the geochemical and data mining techniques together. These methods have been proved successful to deal with the source apportionment issue. In this chapter, the data mining methods used on this topic and implementations in recent years are reviewed. The basic method includes principal component analysis, factor analysis, clustering analysis, positive matrix fractionation, decision tree, Bayesian network, artificial neural network, etc. Source apportionment of trace elements in surface water, ground water, river and lake sediment, soil, air particles, dust are discussed.

Keywords: trace elements, data mining, source apportionment, water, sediment, soil, particles

1. Introduction

On the issue of trace element contamination of environment, the trace elements refer to the elements with lower concentrations than the major elements, O, H, Si, Al, Fe, Ca, Mg, Na, K, Ti, which are usually take no more than 1% in rocks and minerals. The trace elements have attracting wide research attentions for their high potential on environmental contamination and health impact. In some articles, the phase heavy metals are frequently used to represent elements that have high density or is toxic or poisonous at low concentrations. From the view of environmental impact, the phases trace elements and heavy metals refer to similar research objects, which are used as group name for metals and metalloids that have been associated with contamination of water, river sediment, soil and air particles and potential toxicity and ecotoxicity. In this chapter, the phase trace elements (TEs) are used to present the elements that may cause contamination and health problems, and address issue relating to the behavior and mechanism of them among environmental bodies.

The TEs are widely studies in the areas of water, rock, coal geochemistry, leaching and mobility potential, bioaccumulation and human health risk, survey technologies, and other related topics [1, 2]. The harm to human health of TEs are amount related. Some TEs are essential to human in a concentration scale, while become toxic along with the concentration elevation. Some toxic TEs may cause acute and chronic effect even in very low content. In light of the levels of toxicity, trace elements lead (Pb), zinc (Zn), copper (Cu), nickel (Ni), chromium (Cr), cadmium (Cd), arsenic (As), selenium (Se), mercury (Hg), are most investigated, studied and regulated. For example, small amounts of lead in the body can make it difficult for children to learn, pay attention and succeed in school. Lead accounts for most of the cases of pediatric heavy metal poisoning. Arsenic is the most common cause of acute heavy metal poisoning in adults and does not leave the body once it enters. Mercury exposure put newborns at risk of neurological deficits and increased cardiovascular risk in adults.

The TEs may be released from sources of lithogenic or anthropogenic [3, 4]. With the industrialization and urbanization process, TEs released from anthropogenic source are increasing, including discharge of industrial and municipal wastes, storms, run-offs, dry deposition, mine discharge, waste incineration, application of pesticides and fertilizers, sewage irrigation and transportation, and other diffused sources [1, 5–11]. The environmental medias, including water [10, 12–14], sediment, soil [3, 4, 15–23], air particles [15] can be contaminated.

In order to understand and control pollution of the trace element, source identification and quantification of TEs in water, sediment, soil, and particles are of great importance. The traditional techniques are mostly based on geochemical method. Statistical method based on univariate analysis are also used. However, the univariate analysis is cumbersome, and sometimes hard to explain. The multivariate analysis provides a new technique system for the TE source apportionment. Multivariate analysis, and related method, machine learning, data mining have been approved to be successful in a very wide aspects of human living and production. In the area of geochemistry, environmental engineering, applications of the method are also increasingly used.

In this chapter, two related topics are reviewed and discussed. First, the advances of multivariate analysis on the issue of source apportionment, especially several kinds of multivariate analytical method; second, understanding of the contaminating origin of TEs on important environmental media, ground and surface water, sediment in river and lake, soil, precipitate dust, suspended particle matters, PM 2.5 and PM 10.

2. Methods for data mining

To investigate trace element concentration in time series and spatial distribution, migration source, and reaction pathway, technology of data mining is used. In narrow sense, the data mining refers to using multivariate analysis and machine learning method to find distributing or changing pattern in big data sets. In a broader concept, the data mining may include more techniques, such as geochemical, isotopic, univariate analysis, etc. In this chapter, techniques of multivariate analysis and machine learning are emphasized for its increasingly application and effective in source apportionment and reaction path analysis.

Table 1 lists application of data mining methods and implementation on the trace element migration. In which, PCA stands for principal component analysis,

Environmental media	Country/region	Anthropogenic source TEs	Data mining method	References
Water (surface)	Ethiopia	Cu, K	PCA/FA/CA/DA	[24]
Water (surface)	Turkey	Not found	PCA/HCA	[25]
Water (surface)	Belgium	N	Bayesian network	[26]
Water (surface)	USA	Mg, Cl, Na	CA/DA	[27]
Water (surface ground)	Greece	EC, Cl, SO_4, Mg, Ca, Na, K, As, Fe, B, Br, Sr, V (sea water intrusion)	PCA/DA	[28–30]
Water (surface ground)	Spain	—	PCA/regression	[31]
Water (ground)	China	—	PCA/DA	[32]
Water (ground)	China	—	DA	[33]
Water (ground)	China	Se, As, Hg, Cr, Pb	PCA	[34]
Water (ground)	India	Pb, Cu, Cr	PCA/CA	[35]
Water (ground)	India	As, Cd, Co, Pb, V	PCA/CA	[36]
Water (ground)	Greece	N	Bayesian network	[29]
Water (ground)	USA	Cr, Br, Cl, N, S	Semi-supervised ML	[37]
Water (ground)	China	—	Decision tree/CA	[38]
Water/soil	Nigeria	Cd, Cr, Pb, Ni, V	PCA/CA	[12]
Sediment (river)	China	Ni, Hg, Cr, Cu, Cd, Pb, Zn, As	PCA/DA/Monte Carlo	[39]
Sediment (river)	China	Cd, Zn	PCA	[40]
Sediment (river)	China	Cr, Cd, Pb, Hg	PCA, EF	[41]
Sediment (lake)	China	As, Cd, Hg, Pb, Zn	PCA/EF	[42]
Sediment (lake)	China	Cr, Pb, Zn, Cu, Co	PCA	[43]
Soil	China	Hg, Cr, Ni, Ba	PCA/CA	[44]
Soil	China	Cu, Zn, Cd, Hg	PCA	[45]
Soil	China	Not found	PCA/CA	[46]
Soil (peat)	Spain	Cd, Pb, P, Zn	PCA/CA	[47]

Environmental media	Country/region	Anthropogenic source TEs	Data mining method	References
Soil (dust)	China	Zn, Mn, Ni, As, Cu, Pb, Cr, Co	PCA	[48]
Soil (dust)	China	Pb, Cd	PCA/ANN	[49]
Soil	India	Ni, Co	PCA	[50]
Soil (city topsoil)	Armenia	Pb, Zn, Cu, Mo	PCA/CA	[51]
Soil (atmospheric deposition)	China	As, Hg, Cu, Cd, Mo, S, Zn, Cr, Ni, Pb, Se	PCA/CA	[52]
Soil	Spain	Pb, Tl, As, Sb, Cd, Cr, Ni, Be, V, Co	FA	[53]
Soil	Pakistan	Ni, Cr, Zn, Cu, Pb, Cd, Co	PCA/FA/ CA	[54]
Soil (agriculture)	Greece	Cu, Pb, Zn, As, Cd, P, K	PCA/CA	[55]
Soil	Italy	Ni, Cr, Pb, Zn	PCA/FA-MLR	[56]
Soil	China	Cd, Hg, Pb, Zn	PCA/CA	[57]
Soil	Iran	—	Semi-supervised ML	[58]
Soil	USA	—	(Six models)	[59]
Particle	India	Ni, Cu, Pb, Cd, Cr	PCA/CA	[60]
Particle (PM 2.5)	Canada	—	PMF	[61]
Particle (PM 2.5)	China	Cr, Mn, Fe, Cu, Zn, As, Pb, Ba	Regression/Monte Carlo	[62]
Particle (PM 2.5)	China	—	PCA-MLR	[63]
Particle (PM 2.5)	China	Zn, Pb, Cd, Cu	PCA	[64]
Particle (PM 2.5)	USA	—	PCA	[65]
Particle (PM 2.5, PM 10)	Costa Rica	Fe, Ni, V	PMF	[66]
Particle (PM 2.5, PM 10)	Nigeria	Cl, K, V, Cr, Ni, Br, Pb, S, Na, S, Zn, As	PMF	[67]
Particle (PM 2.5, PM 10)	USA	Zn, Pb, P	PMF	[68]

Table 1.
A list of data mining method on trace element migration in recent years.

CA stands for clustering analysis, ANN stands for artificial neural network, FA stand for factor analysis, MLR stands for multi-linear regression, DA stands for discriminate analysis, EF stands for enrichment factor, PMF stands for positive matrix fractionation.

2.1 Geochemical methods

To analyze geochemical properties, and reaction mechanisms, mass balance, piper diagrams, Gibbs diagrams [28] are usually applied. The piper diagram shows the major element composition of water, which category water into different types. Several software, PHREEQC, MINTEQ, geochemists' workbench, can be used to calculate mass balance, saturated index, and model the reaction path, draw piper diagram, etc. By the process of water-rock interaction, major elements and TEs may be released and immigrate to other water bodies, therefore the major elements and trace element with distinguishing feature could be used as source apportionment [69, 70]. However, the TEs undergo geochemical process of adsorption, desorption, mineralization, dissolution to change concentration in water. Therefore, the univariate analysis is not credible and robust. Comparatively, isotopic analysis, both stable and radiogenic [30], may be used as a univariate analysis method or combing some other indexes. The widely used isotopic method are $\delta^{18}O$ and δD in water, $^{87}Sr/^{86}Sr$, $\delta^{34}S$ and $\delta^{18}O$ in sulfate [71, 72], $\delta^{15}N$ and $\delta^{18}O$ in nitrate [29], etc. The isotope $\delta^{18}O$ and δD in water are used to identify water relations between precipitation and surface/ground water. The ratio of strontium isotope of water is strictly controlled by water-rock interaction. For a unisource water, the $^{87}Sr/^{86}Sr$ reflect the mineralogy of the rocks with which the water has been contact and does not change along the water flow. It is highlighted that differences in the strontium isotope ratio and strontium concentration are caused mixing of water of various origins with specific chemical characteristics and isotopic values. Therefore, the strontium isotope is an ideal tracer for element resources, groundwater movements, and water-rock interaction [70, 73–75]. The use of $\delta^{34}S$ and $\delta^{18}O$ in sulfate is increasing because they have wide range of stable isotope composition and the $\delta^{34}S$ value is derived from multiple sources and very close to that of the precursor sulfide mineral. A common anthropogenic source of sulfate is the coal and metal mining which is rich in pyrite and other sulfide minerals. In activity and abandon coal mines, the sulfide mineral may be oxidized and dissolved, with the release of trace elements. It has been proved by the sulfate isotopes that the ground water could be contaminated by the water-rock interaction in coal mines. Besides of the natural isotope, some tracers such as isotopes and stable organic compound are injected into groundwater to find out the flow pathway [71, 72, 76, 77].

To evaluate contamination of TEs and find the source of pollution on water and solid matrix, some calculations are used. The enrichment factor (*EF*) is an enrichment level of a certain TEs in environment, with an equation as shown in Eq. (1):

$$EF = \left(c_i/c_{ref}\right)_{samples} / \left(B_i/B_{ref}\right)_{baseline} \tag{1}$$

where the c_i is the measured concentrations of TEs in samples, the c_{ref} is the measured concentration of the reference element, B_i and B_{ref} are the background level of the local region and reference element in the same region [41]. An *EF* value close to 1 suggests a weathering origin of trace element, while a higher than 1 value means TEs enrichment in soil which is probably caused by human activities. An *EF* value between 2 and 5 indicate a moderate contamination, and a higher than 5 value show a heavily polluted by TEs [49]. The *EF* factor is frequently used combing wit data mining method, or as a verification to trace the contaminating sources.

The geo-cumulative index method (I_{geo}) is defined using Eq. (2):

$$I_{geo} = \log_2[c_i/1.5B_i] \tag{2}$$

where the c_i is the measured concentration of TEs, and the B_i is the background concentration of the particular TEs [41].

The Hakanson potential ecology risk method (RI) was proposed by Hakanson and can be used to evaluate the potential ecological risk posed by TEs in water and solid matrix. This comprehensive method considers four factors: concentration, type of pollutant, toxicity level, and the sensitivity of the water body to metal contamination in water and solid matrix [78, 79].

2.2 Machine learning

Studies of environmental processes exhibit spatial variation within data sets. The ability to derive predictions of risk from field data is a critical path forward in understanding the data and applying the information to land and resource management. Multivariate analysis, or machine learning methods present advantages of precise, robust, and can look insight the phenomena to find mechanism. On the other hand, the environment data usually composed of matrices. Therefore, the machine learning methods is an ideal tool to deal with environmental and geochemical issues. However, the calculation of machine learning and multivariate analysis are complex, which may prohibit their implementation. Thanks to recent advances in predictive modeling, open source software (R, Python, SPSS, SAS, Minitab, etc.), and computing, the power to do this is within grasp.

Basic principle of ML is to train models for the specific data frame using the obtained data, then apply the models on the target problems. The ML methods can be divided into three types, namely supervised, unsupervised and semi-supervised learning. When the training data has labeled data, it is a supervised ML, while unsupervised ML have no labeled data. The semi-supervised ML add labels to data during model training. Generally, the supervised ML has higher precise and robust than others. However, the geochemical and environmental data are usually unlabeled. For example, when the researchers try to identify source of water, or TEs in water and solid matrices, the results are usually not assured. Therefore, the unsupervised ML, up to present, has more widely implemented than the supervised and semi-supervised ML.

2.2.1 Unsupervised ML

It is undoubted that the unsupervised ML is mostly used in this area. Common techniques of unsupervised ML include: principal component analysis (PCA), factor analysis (FA), clustering analysis (CA), positive matrix factorization (PMF), etc.

In the scope of machine learning algorithm, PCA is a tool to reduce high dimensional matrix to a lower, usually two to five, dimensional matrix. Dimensional reduction is accomplished by transforming the data to a new set of variables (principal components), which are derived from linear combinations of the original variables and classified in such a way that the first principal components are responsible for most of the variation in all of the original variables [80].

A matrix M, with m observations (row) and n variates (column), is calculated following five steps to form a new matrix with less variates.

Step 1: the raw data in the M is standardized;

Step 2: the covariance matrix of the standardized M is calculated;

Step 3: eigen value and eigen vectors of the covariance matrix are calculated;

Step 4: contributing ratio and accumulative contribution of the eigen value was calculated, then the principal components can be determined according to mathematical and project criterion;

Step 5: loading of every principal component and score of every observation can be calculated.

Theoretically, the number of new variates is equal to variates of the original matrix. On the other hand, the new variates contribute different ratio to explain variance of variates, then the principal components are selected based on the explanation ratio. Different criterion was used, some researchers use the eigen value larger than 1, and some others use the accumulative contribution of the eigen values, say 80%.

Given the M has 30 observations and 10 variates (*x1*, *x2*, …, *x10*), and three principal components (*y1*, *y2*, *y3*) are selected, eigen vector of *y1*, *y2*, and *y3* are *A1*, *A2*, and *A3*, respectively, then:

$y1 = A1 * x1$
$y2 = A2 * x2$, and
$y3 = A3 * x3$.

After the calculation of PCA, some variates have higher loadings on specific principal components, while some variates have higher loadings on other PCs. Then it is inferred that the variates have similar pattern in the matrix may have similar pattern in the real world, i.e., the source, migration behavior, and reaction pathway. Theoretically, the PCA is similar with clustering analysis, but the PCA is not constrained to two dimensions, which allow the researchers mine the inner relationships in the matrix and understand real world more precise.

The Factor analysis (FA) is based on PCA, have similar principle, and aim to obtain similar result with PCA, but the applications are less than PCA.

The data on every variate should be normal distributed. Kaiser-Meyer-Olkin (KMO) test and Bartlett's sphericity test are usually used to determine the distribution of data for analysis of PCA/FA.

Because of the advantages of PCA and FA in analysis of environmental and geochemical data mining, they are widely implemented [1–4, 11, 12, 17, 19, 23, 28–30, 34, 41, 49, 64, 78, 81–83].

On the issue of source apportionment of particle matter and trace elements for the suspended particles and trace elements inside, a method of positive matrix factorization (PMF) is usually used. When we have a matrix *M*, with *f* of observations, and *n* of variates, the *M* can be calculated as Eq. (3):

$$M_{(f*n)=}W_{\mathrm{f}*k} * H_{k*n} \tag{3}$$

In a source apportionment problem, *W* stands for source contributions, *H* stands for source profiles, *k* is the number of possibly sources. The least loss function determines the proper *k* value and the matrix *W* and *K*, then the source quantity, contribution ratio can be inferred. The W_{f*k} need to be normalized by their average value across all samples as shown in Eq. (4):

$$\overline{w}_{fk} = \sum_{f=1}^{n} \frac{w_{fk}}{n} \tag{4}$$

where $\overline{w}_{fk}$ is elements in the matrix *W*. The PMF is usually used in source apportionment for particle, such as PM 10 and PM 2.5 [61, 66–68], but seldom used in other environmental medias.

2.2.2 Bayesian network

A model of Bayesian Network has been implemented to estimate TE source contribution, and evaluate the contaminating levels [26, 29, 84–86]. A R package SIAR (Stable Isotope Analysis in R) can be run to calculate the isotope mixing model base on the Bayesian Network. The mixing model can be elucidated as the equation set Eq. (5):

$$\begin{aligned} X_{ij} &= \sum_{k=1}^{k} p_k \left(S_{jk} + c_{jk}\right) + \varepsilon_{ij} \\ S_{jk} &\sim N\left(\mu_{jk}, \omega_{jk}^2\right) \\ c_{jk} &\sim N\left(\lambda_{jk}, \tau_{jk}^2\right) \\ \varepsilon_{jk} &\sim N\left(0, \sigma_j^2\right) \end{aligned} \tag{5}$$

where X_{ij} is the isotope value j of the mixture i, in which i = 1, 2, 3, ..., N and j = 1, 2, 3, ..., J; S_{jk} is the source value k on isotope j(k = 1, 2, 3, ..., K), c_{jk} is the isotope fractionation factor for isotope j on source k. p_k is the proportion of source k, which needs to be estimated by the SIAR model. The S_{jk} and c_{jk} are normally distributed with mean μ_{jk} and standard deviation ω_{jk}, mean λ_{jk} and standard deviation τ_{jk}, respectively. ε_{ij} is the residual error representing the additional unquantified variation between individual mixtures and is normally distributed with mean 0 and standard deviation σ_j. Algorithm of Monte Carlo is usually to solve the equation of Bayesian network.

2.2.3 Decision tree

The decision tree is a kind of supervised ML, including a series of machine learning techniques to divide samples into different categories, such as algorithms ID 3, C 4.5, C 5.0, CART, etc. Different algorithms follow the same principle: the observations are divided by breakpoints on a variate. The selection of variates and breakpoints provide basis of decision, and all of the decisions make a tree for users to make a project decision system. Take the algorithm CART for example, the sample space needs to be split by using the variate breakpoints. Different split strategies make different decision efficiency. As an index, the Gini Coefficients are use. The decision tree machine will calculate Gini index for every split, the split method with lower Gini Coefficients is used. Advantages of the decision tree are easy to carry out and easy to explain. Some researchers have introduced the method to trace source of nitrogen and TE contaminations [38].

Some decision tree series methods, including random forest, boosting method are widely used in the data mining [59]. However, the applications in the source apportionment are rare. The obstacle that prohibit the implementation of the decision tree methods may be the acquirement of data labeled.

2.2.4 Artificial neural network

The artificial neural network (ANN) has been recognized as a powerful supervised ML and applied in a wide scope of engineering and research. The ANN research is one of the most active research in the ML algorithm, which have a lot of branches. It is also the basis of deep learning, which is used as figure and voice identification.

In the area of environmental and geochemistry, the implementation of ANN is not as popular as unsupervised ML. The most important reason is that it is a kind of supervised data mining method. In the data preparation step, the observations need to be labeled, while the environmental data are usually cannot or difficult to label. However, some researches have used this to predict the contaminating potential, Mclean et al. reviewed the application of ANN on ambient air pollution [87]. The second obstacle for its implementation is that ANN usually need larger amount of data than PCA, decision tree, and some other methods.

A basic ANN model has an input layer, one or several hidden layer, and an output layer. Variates of one observation are input through the input layer, and the output layer are labeled data, the hidden layer are used to calculate the model from input to output. The relationship of the layers is trained by input the variate data and labeled data. A trained model is used to predict while input data are obtained. Once the labeled data can be obtained, the ANN model is useful to predict, discriminate, divide samples with trained model in the engineering and research of geochemical and environmental purpose.

2.2.5 Discrimination analysis

The discrimination analysis (DA) is a kind of supervised ML, because the data set is labeled in the model training step. The DA has a similar concept with the principal component analysis. While the PCA tries to find principal components (new axis) to stand for the most variates, the DA tries to find axis that stand for the least variates, so that the different variates and observations can be divided. Prior to the application of DA, all of the variables were standardized to ensure that scale differences between the variables are eliminated. Hence, the absolute discriminant weights ranked the variables in terms of their discriminating power, i.e., the variables with large weights are those that contribute most to differentiating the groups.

In the model training step, if the origin of samples can be identified, then the DA could be used to identify sample source. For example, water intrusion in coal mines may come from different ground water aquifer, the different aquifers have specific geochemical characteristics and hazard level. As a water hazard control work, discrimination models can be set up by train the labeled data. The labeled data means water collected from different aquifers. Once water intrusion happens, the water characteristics is used to identify source of water by comparing with the model [27, 33].

The criterion applied in the discrimination analysis are mainly distance based or Bayes rule base. When the distance rule is used, Manhattan distance of samples to different groups are calculated, a group with less distance to the samples is labeled to the samples. The distance has to be calculated in pair, which constrains efficiency of the model training and implantation. Another popular method is called the Bayes discrimination method. A reasonable way to discriminate the group of a characteristic sample is to compare the conditional probability of the characteristic sample falling in different category. The class with the highest conditional probability is the final category result of this characteristic sample. Theoretically, the Bayes DA has a higher coefficient and accurate than the distance-based DA.

2.2.6 Semi-supervised machine learning

The unsupervised ML are easy to carry out, but low in accuracy, robust, reliability and duplicate, while reverse for the supervised ML method. As an improved strategy, the semi-supervised ML is an option. When the labeled data is not easy to acquire, and need to do unsupervised ML at first, then the semi-supervised algorithm may apply to add labels for the data while the model is being trained. Vesselinov et al. used the non-negative matrix factorization method for blind source separation in the first step, then a semi-supervised clustering algorithm was used to predict the sources of contaminates [37]. Fatehi and Asadi used a hybrid method combining hieratical clustering and fuzzy c-means clustering to classify soil types [58]. At present, this method used in this topic at present is rare.

2.3 Regression

The regression is easy to use, explain, and understand. Also, the regression is a big too box in the machine learning workshop. The most popular method is the multivariate linear regression, sometime logistical regression, lasso regression, ridge regression, plastic net regression can also be used. The regression method can be combined into other machine learning techniques, such as decision tree [56], support vector machine, etc. However, the regression has very distinction shortages. In a regression process, the model of data is fitted to a linear or curve function, which may not accord with the real situation. Second, the regression is prone to overfitted, while the training mode performed well, disaster results may be gotten when applied in real environment. To solve this problem, lasso, ridge, and plastic net regression are applied. Besides the two issues, another problem may bother the application of the regression, the data are usually not easy, or cannot to label. In this situation, unsupervised techniques should be used. Once the labeled data are acquired, regression method are applied [31, 56, 59].

2.4 Artificial tracers

In order to find ground-surface, ground-ground water relationship, artificial tracers are also used. The chemical traces sodium chloride, eosine, uranine and pyranine were used to analyze spring-ground water relationship. Conductivity meter and thermometer was yet installed for electrical conductivity (EC) monitoring and field fluorimeter was equipped for tracer detection [31].

2.5 Other methods

In a research from Alaska America, six models were set up to predict soil contamination. The model includes random forests, generalized boosted regression, elastic net regression, multivariate adaptive regression splines, generalized linear model with stepwise selection using Akaike's information regression, and partial least squares regression. Although got similar explanatory power overall among the models, the machine learning models performed much better than the linear models on predictive accuracy and were better able to identify variables of interest and describe non-linear relationships. In order to understanding the mechanisms behind trace element pollutant fate and transport and were less vulnerable to errors of omission, the machine learning techniques have priorities than the linear models [59].

3. Implementation of the data mining of TE source apportionment

The environmental medias that may be contaminated by trace elements are grouped into four types, water, sediment, soil, and particles in this chapter. In every case, the probably sources of trace elements are listed in order of importance. The main method to be used are listed in **Table 1**.

3.1 TE apportionment in water

3.1.1 Contamination sources of surface water

The TEs migrate from rock and coal to water through water-rock interaction. Then the surface and ground water may be contaminated.

In Turkey, TEs source in a large reservoir was identified. The PCA showed that PC1, PC2, and PC3 includes Co/Cr/Fe, Cu/Pb/Zn, and As/Cd, respectively. Combing with correlation analysis, the three PCs were identified to natural source, bedrock weathering, and bedrock weathering, respectively [25]. Another research revealed by PCA that mineral pollution, nutrient pollution, and organic pollution are major latent factors which influence the water quality of Asi River [88].

Because of the vandalization of pipeline, soil and water may be contaminated. The PCA results showed that the first source was associated with anthropogenic source, such as vehicular emission, which was composited by Cd, Cr, Pb, and Mn. The second source, including Cu and Zn, was related to natural geological origin, and the Ni and V were released from natural source collaborating with the petroleum contamination [12].

In Ethiopia, water samples were divided into four categories by clustering analysis: natural cluster, mixed cluster, agriculture cluster and urban cluster. In the agriculture cluster, VF1 has strong loadings on TN, NO_3^-, salinity, Fe, NH_3, hardness, and Mn, which is cultivated originated, VF2 were associate with turbidity, Chl-α, and Cu, which may come from farming and excavation sites of quarrying activities. Mg, and K were mainly loading on VF3, and VF4, respectively. K is mainly spread while potash fertilizer is used [24].

Supervised ML technique, discriminant analysis, was applied with the clustering analysis to assort and find spatiotemporal distribution of trace element in surface water, in the USA. Sources of salt ions (magnesium, chloride, and sodium) vary from natural sources (oceans, atmospheric deposition, weathering of common rocks, minerals and soils, and salt deposits and brines) to anthropogenic sources (landfills, wastewater and water treatment, agriculture, and application of deicing salts) [27].

A Bayesian isotope mixing model was used to estimate proportional contributions of multiple nitrate sources in surface water in Belgium. The result showed that "manure and sewage" contributed highest, "soil N", "NO_3^- fertilizer" and "NH_4^+ fertilizer and rain" contributed middle, and "NO_3^- in precipitation" contributed least [26].

3.1.2 Contamination sources of ground water

In southern India, potential TE source of ground water was analyzed, it was concluded that Fe and Mn were natural origin, Cr, Cu, Pb and Ni may come from mixed sources, natural and flow contaminated with fertilizers and pesticide. In another study of northern India, the sources of ground water were identified to be anthropogenic source via agrochemical and industrial wastes (As, Cd, Co, Pb and V), parent material from an adjacent area (U and Sr), lithogenic origin (Fe, Mn, Zn), and background level elements (Mo and Se), respectively [36].

In Greece, Matiatos et al. [28–30] investigated surface water and ground water combing the method of geochemical, isotopic and multivariate statistical analysis, such as PCA and Bayesian isotope mixing model. By the PCA analytical result, EC, Na, K, Cl, and Mg, were found to be seawater inrush origin, Fe/Zn, Ca/hardness were from water-silicate rocks interaction, and dissolution of limestone, respectively, NO_3 stand for nitrogen pollution, and SO_4^{2-} and Mn were from dedolomitization process and increased agricultural input.

A semi-supervised ML technique was used to trace contaminants' source in the USA. Vesselinov et al. [37] proposed a contaminant source identification approach that performed decomposition of the observation mixtures based on non-negative matrix factorization (NMF) method for blind source separation (BSS), coupled with a custom semi-supervised clustering algorithm. As a result, the mixing coefficients of all the groundwater types (contaminant sources) for each observation well (samples) were obtained.

As a supervised ML technique, decision tree is used combing with isotope method in a study to determine nitrogen source in groundwater. The decision tree has made 97.5% success in the water quality analysis. However, concentration data alone could not identify the dominant NO_3^- sources for groundwater contamination. It is suggested that an integrated approach should be setup by the combination of the N and O isotopes of NO_3^- with land-uses and physical-chemical properties, especially in areas with specific activities [38].

3.1.3 Contamination by coal mine water

One of the most focused issues of surface and ground water contamination is the acid mine drainage (AMD). The AMD is formed when pyrite and other sulfide minerals oxidized and dissolved during coal and metal mining, highway construction, and other large-scale excavation [13]. In an anaerobic environment, the sulfide minerals are stable, while exposure to water and oxygen, and with other accelerating factors such as bacteria, they are oxidized to form sulfuric acid, accompanying release of trace elements to surrounding water bodies [89]. In coal mine water, some of the drainage is alkaline, the leaching behavior and the TE composition in the leaching water are different [14].

Mobility of the TEs in AMD depends on several conditions. First, what is the TE occurrence and abundance in the potential AMD source; second, during the water-rock interaction process, where and how the adsorption-desorption, dissolution-precipitation take place; third, what are the main flow path, and the river and lake geochemistry where the TEs may be adsorbed or released again.

If the flow path is known, the source and reaction rates of specific trace elements can be estimated by mass balance calculation. The post-dissolution behavior of TEs is controlled by solution composition, pH, Eh of the water, temperature, and contact-time with mineral surfaces. For example, metal elements will have little attenuation in the solid phases, and high mobility potential into water. The versus behavior can be observed for the metalloid elements. Along with the flow path, water geochemical characteristics and pH and Eh of water changes, the TEs may undergo very complex reaction process, leading to redistribution of TEs in surface water, ground water, and sediment in the water bodies. Therefore, the source identification of TEs is an important and challenging work.

The pH of AMD ranges from 2 to 8. In an acid environment, metal element, Pb, Cd, Cu, Ni, have high mobility, while some metalloid element, As, Se, tend to migrate in an alkaline environment. Damaging effects of AMD are reported in Asia [34, 90–93], North America [72, 94–96], Europe [83, 97], South America [47, 98, 99]. When AMD enters surface water bodies, the effects include biotic impacts on stream and lake organisms through direct toxicity, habitat alteration by metal precipitates, visual changes from orange or yellow staining of stream sediments, nutrient cycle disruptions, or other mechanisms, and the water often becomes unsuitable for domestic, agricultural, and industrial uses. Gammons et al. have found the contamination of abandon coal mines on ground water using method of isotope analysis [71, 72].

The TEs in coal are not only migrate while mining and dumping of gangue and dust deposit [100], but also accompanying spread by smoke, fly ash, bottom ash when combustion [101, 102]. Trace elements, As, Cu, Se are found concentrated in the fly ash, which indicate impact on water and soil quality [103, 104].

3.1.4 Source apportionment of water inrush in coal mines

The TE source apportionment technology is used in coal mines to determine the source of water inrush [32]. In coal mine, water inrush constantly threatens the production, human health and cause financial losses. The water inrushes are cauterized to four sources: quaternary sand-gravel pore aquifer, Dyas sandstone aquifer, limestone aquifer from Ordovician and Carboniferous, and abandoned coal mine districts, respectively. Different sources show varies features and need different treatment strategies. The main purpose of the water inrush analysis is to find categories of source aquifers. Huang et al. [32] proposed a technology system, Piper-PCA-Bayes-LOOCV discrimination model to determine water inrush types in coal mines. The piper diagram is a geochemical technique to show the water characteristics, and abnormal samples/points were screened in this research. PCA was used to lower dimension of the sample matrix, to make less variates standing for all the original variates. Then the supervised ML model, Bayes DA, is used to train and implement a model for water source discriminant. LOOCV means leave-one-out cross-validation, to validate and improve quality of the model. Wang et al. used discriminant analysis to determine water bursting source in coal mines [33].

3.1.5 TE occurrence and reaction pathway

The PCA method has also used to investigate trace element occurrence in rock/coal, and reaction pathway, which may be the source of TEs that have contaminating potential on surrounding water bodies. Shan et al. [34] found that in coal host rock seam, Se/Cd/Hg/ As occurred in sulfide minerals, Be and V occurred in carbonate minerals, Cr and Pb occurred in clay minerals, respectively; while in coal seam, Se/Cr/Pb occurred in clay minerals, As and Hg occurred in sulfide minerals. Se, As and Hg immigrated through dissolution of sulfide minerals, Cr immigrated through transformation of clay minerals in coal host rock. In coal seam, As and Hg occurred in sulfide minerals. Se, Pb and Cr immigrated through transformation of clay minerals, As and Hg immigrated through dissolution of sulfide minerals, respectively. Pumure et al. [105] investigated occurrence of selenium and arsenic in coal by the method of two step PCA, founding that ultrasound leachable selenium concentrations were associated with 14 Å d-spacing phyllosilicate clays (chlorite, montmorillonite and vermiculite all 2:1 layered clays) whilst ultrasound leachable arsenic concentrations were closely related to the concentration of illite, another 2:1 phyllosilicate clay.

3.2 TE apportionment in sediment

Surface water and sediment compose a reaction system, trace elements in water may be adsorbed by sediment, meanwhile, trace elements in sediment are released. Therefore, the sediment may be a sink or origin of trace elements. Because of the complex reaction pathway, and environmental persistence and biological accumulation, the trace elements in the aquatic environments has drawn special attentions [106].

In southwest China, lake sediment was analyzed [42]. PCA result showed that Cd/Hg/Pb/Zn, and As (as PC2 and PC3, respectively) were mainly from non-point anthropogenic sources, especially with the atmospheric emission from non-ferrous metal smelting and coal consumption [107].

In Jiangxi China, river sediment was investigated. As the metal mines are excavating in the study area, metal element contamination was found. The PCA analytical result show probably coal and gold mining, copper mining and refining, Zn/Pb deposits and agricultural activities origin associated with PC1, PC2, and PC3, respectively. The PC1 were high loaded with Ni, Hg, Cr, the PC2 were high loaded with Cu, the PC3 were high loaded with Cd, Pb, Zn, As, respectively [39]. A research on lake sediment in Jiangxi China showed that Cr, Pb, and Zn may be mainly derived from both lithogenic and human activities, such as atmospheric and river inflow transportation, whereas Cu and Cd may be mainly contributed from anthropogenic sources, such as mining activities and fertilizer application [43]. In northern China, Cd and Zn are found originating from agriculture source and Cu, Cr, Ni were natural source origin [40]. In northwest China, Zn, Cu, Ni, and As were high loaded on the PC1, and natural originated, Cr and Cd/Pb/Hg are high loaded on the PC2 and PC3, which were township/silicon chemical factories, and agriculture/urban construction origin, respectively [41].

3.3 TE apportionment in soil

Researches have focused on distinguish TE source from natural and anthropogenic [108, 109] contaminates in soil. The major natural contribution of heavy metals comes from the parent materials from which the soils developed. The anthropogenic source of heavy metals in soils includes acid mine drainage [110], agricultural and industrial waste discharges [111], atmospheric deposition [112], fertilizers and pesticides [113], which has a significant contribution to the content levels of heavy elements in soils. PCA are now a popular technique to trace source of TEs in soil, then enrichment factors are usually used to verify the sources. In order to investigate source of TEs and spatial distribution, the combination method of geochemical, multivariate analysis, and geostatistical analysis. GIS and multivariate analysis of soil contamination has been detailed reviewed [114]. Understanding sources of heavy metals in surface soils is imperative for the decision to implement the strategies for protecting the food safety, human health and ecosystem sustainability.

3.3.1 TE apportionment in agricultural soil

In Greece, two main sources explained 74.8% of all the variance for the agriculture soil contamination analysis. The TEs Cu, Pb, Zn, As, Cd, P and K were identified to be anthropogenic influence, and TEs Ni, Co, Fe and Cr were recognized to be natural source origin [55].

In soil samples on hills in India, four principal components were determining by using the PCA method, high loading TEs on which are Mn/Zn, Cr, Ni, Co, respectively. The PC1 and PC2 were inferred to be natural sources, and PC3 represent fossil fuel burning origin, which contribute most of Ni in soil, and PC4 represent irrigation sources, respectively [50].

In Shanxi province China, soil samples were collected in an area of 25k km^2. The PCA analytical result showed that Co, Cr, Cu, Mn, Ni, Se, V, and Zn were mainly originated from natural source, and Cd and Pb were affected by anthropogenic pollution heavily. Associated with the spatial data, Pb were strongly associated with road traffic, and Cd were linked to industrial activities. In order to predict Pb and Cd concentration in the following years, an ANN model was applied [49].

In Beijing China, the TEs can be represented by two PCs. The TEs Co, Ni, Cr and V were probably released from parent material of the soil, Cd, Cu, and Zn were primarily from agricultural cultivation. Hg may be originated from coal combustion or mineral fertilizers [45].

In Jijin China, Al, Fe, Mn, Zn, Cr, Ni, As, Cu, and Pb was found accounting for 55.16% of the total variance, which was identified as natural source. N, OC, P, Cd, and Hg have high loadings on the PC1, accounting for 16.75% of the total variance. The PC2 also seemed as natural source, high relationship of Hg and Cd was explained to high organic affinity [46].

In Iran, a kind of semi-supervised ML method was applied. The study area was located around a Cu-Au porphyry deposit, so the soil may be associated. Initially eleven soil geochemical variables were selected by using hieratical clustering analysis and expert knowledge. Then, the semi-supervised fuzzy c-means clustering method (ssFCM) was used to separate multivariate soil geochemical anomalies from back-ground for further drilling [58].

3.3.2 TE apportionment in urban and industrial top soil

The impact of ore deposit on surrounding soil was investigated in Beijing China. Frequent mining activities produce dust, acidic drainage from the oxides and mill tailing. Cu, Co, Zn, Cd and V was found to mixed sources originated; Be, Pb and As came from natural sources and are mainly affected by the weathering and erosion of parent rock material; Cr, Ni and Ba were polluted by fine particle, industrial and mining activities; transportation and soil minerals were the common sources of Cr and Ni; Hg came from anthropogenic sources, mainly impacted by mining, beneficiation, smelting and acid mine drainage waste [44].

Large urban and industrial areas along the coastline in Italy was investigated. Pb and Zn due to heavy traffic and alloy production. Some Cr and Ni contamination were discerned through releases from tannery industry. Zn and Pb enrichment were mainly related to the large volcanic complexes. Cr and Ni were enriched in the siliciclastic deposits [56]. Another large-scale investigation was carried out in Yangtze river delta China, industrialization lead to high contamination potential on environment. Four PCs were selected to present the sources of trace element. As, Hg, Cu, Cd, Mo, S and Zn are recognized as traffic origin. Fe and Mn were from natural resources. The PC3, including Cr and Ni, pointed to pyrometallurgical processes, especially non-ferrous metal industries, etc. The PC4, composited by Pb and Se, was inferred to be coal combustion originated [52]. Another research carried out in this area showed that Cr, Ni, Co, Mn, Cu, and As were mainly came from natural sources. Cd/Hg and Pb/Zn originated from anthropogenic sources in two different groups [57].

In Shaanxi province China, roadway dust was analyzed. TEs Zn, Mn, Ni, As had the highest variance. Because Zn was released mainly from wear vehicle tire and corrosion of galvanized automobile part. Cu, Pb, and Cr was inferred to traffic origin. The third source was dominated by Co and Ni, and they were released from machine manufacturing plant [48]. In a research from Alaska America, soil contamination was found to be caused and controlled mainly by distance to road, traffic category, including highway and refuge road, land cover category, paved or not, land cover category, traffic loading, and other parameters, in descending order [59].

In Pakistan, four factors were identified using the factor analysis to trace surface soil contamination in industrial cite. VF1 contains Ni, Cr, Zn, and Cu, which originate from vehicular emission and industrial activities. VF2, compositing by Pb, Cd, and Co, originated from anthropogenic activities such as automobiles. Fe, Mn, standing for VF3, and VF4, were natural source origin [54]. In Armenia, Ti, V, Mn, Fe and Co, were identified to be natural originated. The PC2 include two distinguished negative groups, As/Hg, and Pb/Zn. The PC3 is composited mainly with Cu and Mo, and recognized as anthropogenic origin [51]. In Spain, the first source, including Pb, Tl, As, Sb, Cd, pointed to coal combustion. The second source was traffic air pollution origin, which released Cr, Ni, Be, V, Co. The third and fourth factors explained a very low proportion of variance and were considered secondary. These factors included TEs Cu, Zn and Sn, showing mixed behavior with regard to the first two factors [53].

In Nigeria, because of the vandalization of pipeline, the soil and water may be contaminated. The PCA gave 78.68% of accumulative contribution of the covariance from the first three PCs. PCA analysis result in soil was similar with that in water [12].

3.3.3 TE apportionment in soil to recall human activities

In Spain, core was obtained from peat bog, to evaluate trace element distribution and human activity impact in the past 8000 years. It was found that Al, Ba, Cr, Ga, K, Na, Sr, Ti, V, Y and Zr were lithogenic and supplied by atmospheric soil dust, while Cd, Pb, P, and Zn were recognized to anthropogenic, especially the ore exploration. The depth of samples depicted the influence degree of human activities yearly [47]. The EF profile showed that Pb, Zn, and Hg were at peak values in atmospheric in Roman age and nineteenth to twentieth centuries.

From the recent researches, it is concluded that Pb is an important anthropogenic originated element. Some reports argued that the vehicle emissions, brake lining, coal burning, plastics and rubber production, and car barriers are potential source of Pb. Meanwhile, Cu might come from vehicle brake lining, Zn from vehicle tires [51]. For the agriculture soil, Cu are usually cumulated by application of commercial fertilizers and Cu-based pesticides and fungicides [115]. Cd was related to the use of phosphate fertilizers [116]. Mineral fertilizers and animal manure may lead to elevation of Zn and Cu levels in soil.

3.4 TE apportionment in air and particles

The TEs spread through the air usually as particles. The particle smaller than 10 μm is called PM 10, while PM 2.5 stand for that smaller than 2.5 μm. It is obviously that the haze-day rate has increasing in the past decade, several researchers have reported characteristics, composition, and sources of PM 10 and PM 2.5 in some Chinese cities [17, 64, 117]. At the same time, PM 10 and PM 2.5 in megacities all around the world are investigated [118–121]. TEs, such as Cu, Zn, Pb, Cd, Cr, relating to the PM 2.5 and PM 10, show deleterious effects to human health. Based on the epidemiological and toxicological studies [122, 123], the TEs in ambient PM 2.5 influence the severity of allergic respiratory disease and have a high cancer risk to the exposed populations [81, 82].

In the source apportionment analysis, six types of main resource of ambient particular matter are commonly found: natural sources (including soil dust and sea salt), domestic fuel burning, industry, traffic, unspecified source of human origin pollution. Soil dust refer to the bare soils by local wind. Sea salt particles can be found close to the coast. Domestic fuel burning includes coal, gas fuel and wood for cooking and heating. Traffic is a complex source of PM and TEs. All the burning of fuel and diesel, wear of brake linings, clutch, and tires are source of TEs [124]. The

"Unspecified sources of human origin" category mainly includes secondary particles formed from unspecified pollution sources of human origin. The reasons of the second outbreak of PM 2.5 are complex, including some chemical reaction. In fact, the reasons of the fog and haze are: (1) the accumulation of the fog and haze, namely the results of combustion, automobile exhaust, and dust effects; (2) the fog and haze particles' upward momentum—hot-air upward movement and wireless communication, namely the electromagnetic wave net sports; (3) no sustained wind. These three conditions indispensable lead to persistent fog and haze weather, and the second outbreak of PM 2.5 results from the above three conditions together. In a review for the source apportionment study, 87% of the record have traffic origin, 66% have industry origin, 45, 100, and 89% have domestic fuel burning, unspecified source of human origin, and natural sources origin, respectively [125].

In southern China, TE source in the PM 2.5 was identified using PCA technique. Three sampling sites were analyzed separately. In YL sampling site, the PC1, with Zn and Pb, were identified as the traffic source, Cu and Cd, high loading on the PC2, originated from coal and other kind of fossil fuel. In KF sampling site, Zn, Cd, and Pb were from vehicle emission and abrasion of automobile tire. Cu, high loaded on the PC2, is a tracer of fossil and other fuel combustion. In the YH site, Cu, Cd, and Pb were associated with domestic fossil fuel burning, and Zn represent brake and tire wear and other transportation processes [17]. During Chinese Spring Festival, haze may occur more frequently, and the PM 2.5 level can be elevated. In Henan province China, sources of PM 2.5 were identified by using PCA, and a model to predict PM 2.5 concentrations using multivariate linear regression was set up. The most important source was burning source, including coal combustion, fireworks, fire crackers and biomass burning, contributing 61% of all the PM 2.5. The second, third, and fourth sources were vehicle emission (27%), soil (8%), and road dust (3.28%), respectively [63].

In Costa Rica, by using the method of PMF, eight important sources of PM 2.5 and PM 10 and TEs were identified. Vehicle exhaust, containing EC, OC, SO_4^{2-} and certain amount of Fe, residual oil combustion, bringing Ni and V, fresh sea salt, including Cl^-, Na and Mg, were the first three source. The others are crustal, or dust aerosols originated, organic carbon and sulfate, secondary sulfate, secondary nitrate, and heavy fuels [66].

In the USA, sources of PM 2.5 were determined, variance of meat, secondary aerosols, motor oil/brake dust/other outdoor, dust, cigarette, gasoline, biomass burning, and retene were explained with 23, 14, 10, 9, 6, 6, 5, and 5%, respectively. For the chemicals, meat released cholesterol, the alkanoic acids, OC, and light n-alkanes. Ammonium, sulfur, and nitrate were mainly released by secondary aerosol. The PC3 includes cholestanes, hopanes, Ba, and nitrate, which was related to motor oil/brake dust/other outdoor [65].

A 6-year investigation of PM 2.5 levels, source and potential human risk was investigated in Canada. Secondary organic aerosol, secondary nitrate, secondary sulfate, transportation and biomass burning, contributed more than 85% to PM 2.5, the importance of which was in descent order [61].

In Nigeria, source of PM 2.5 was identified to be soil (44%), savannah burning (26%), scrap processing (18%) and vehicular emissions (12%), and soil plus biomass burning (71%), sea salt (22%), scrap processing (5%) and vehicle emissions (tire wear) (2%) for the PM 10. Elements Al, Si, Ca, Ti, Mg, Fe and Na were spread through fine particle, and crustal elements Al, Si, Ca, Ti, Mn, Fe and anthropogenic elements Cl, K, V, Cr, Ni, Br, Pb, and black carbon were spread through coarse particles. Savannah burning release Br, BC and Pb through fine particles. The vehicle emit Na, S, Zn, As, Br, and Pb, and spread through fine particles [67].

Coal mining impact of air pollution, including suspended particles was investigated in India. The PCA and CA results suggested PC1 represent PM 10, SO_2, PM 2.5, PM 1.0, Ni and Cu, which are originate from coal burning and active mine fire. PC2 was high loaded with NO_2, Pb, Cd and Cr, and originated from crude oil combustion and vehicular emission. The PC3, including Fe and Mn, was mainly contributed by earth crust, wind-blown soil, and coal fly ash [60].

In the USA, brake wear, tire wear, fertilized soil, and resuspended soil were found to be important sources of copper, zinc, phosphorus, and silicon, respectively, using the method of positive matrix factorization. Zn was found strongly related to tire wear but also contributed to the Pb-rich features and soil. At the same time, the Pb-rich contributions are highly correlated with the tire wear, elevated P contributions within the fertilized soil as well as the Pb-rich feature [68].

Brinkman et al. compared the performance of PCA and PMF on the source apportionment for the particle matters. It was found that most of the PCA factors were easily distinguishable from others by sharp differences in the factor loadings. For many individual compounds, the variance was explained primarily by a single factor. In contrast, the factors obtained with PMF were more difficult to distinguish because anticipated tracer compounds for certain sources appeared in multiple PMF factors [65].

3.5 Summary of method used to identify source of contaminates

Applications and implementations of multivariate analysis/data mining, combining with geochemical method, on source apportionment of trace element as contaminates in environmental medias are increasing, with the development of techniques of big data, machine learning, and computer software. Four environmental medias, water, sediment, soil, and particles are discussed.

Four types of application can be identified for water contamination: trace the source of TEs, evaluate water quality of surface water and ground water, identify intrusion in coal mines and other scenario, and find and quantify water relationship between different bodies, such as surface-ground water relationship. The sediment and water composite a reaction system, i.e., the sediment could be origin, sink of trace elements in water, or be sink at first step, then origin again. Therefore, the system should be analyzed together. The researches on sediment are less than water, and most of articles on this topic are from China.

The most used method for the source apportionment of TEs in water and sediment is principal component analysis (PCA), probably for it's easy to use and explain. With the developing of data mining algorithm and calculation software, the application of PCA become easier and more efficient. The similar method, factor analysis (FA) is also used. The PCA and FA are both unsupervised ML method. Although having less accuracy than the supervised method, these methods are suitable for this topic.

Supervised ML methods are also used in this area, though much less than the unsupervised ML methods, and its scope of application is different. For example, decision tree is used to classify the sample types [38]. Discriminant analysis is also a supervised method, its implementation can be found, especially on the identifying water inrush source in coal mines, as the labeled data can be obtained [32, 33]. In this sense, other supervised machine learning method, ANN, support vector machine, decision tree, can also be used to identify water inrush source. Usually, ANN need more data to improve predicting quality, than SVM and decision tree.

In order to combing the advantages of unsupervised and supervised machine learning methods, semi-supervised method has been introduced and implemented on this topic [52]. At present, related researches are rare, but promising reports are expected.

From the reviewed reports, it is concluded that the surface water is more contaminated by major elements, and nitrogen, which may stand for the organic contamination. The ground water is more contaminated by trace elements, As, Cr, Cd, Pb, Hg, Se, etc. The surface water may be impacted by civil and industrial activities, and the ground water may be impacted by water-rock interaction in the rock seam. The most important anthropogenic source of trace elements in the ground water is the coal and metal mines. These mines contain high content of toxic trace elements, which is stable in an anaerobic environment. Once the rock and coal are excavated, trace elements are released. Less contaminated by trace elements in the investigated surface water is not proving of safety of the surface water. Researches of sediment in rivers and lakes have found high content of anthropogenic source trace element, including As, Cr, Cd, Pb, Hg, Se, Cu, Zn, Ni, etc. The sediment and water in river and lake composite a reactive system, in which the sediment is both sink and source of the trace elements. Therefore, the source, reaction pathway in this system need thoroughly researches and regulations.

Researches on soil can roughly be divided into two large group, agriculture soil, and urban/industrial soil. Unsurprisingly, first TE source of agriculture soil is natural, and first TE source of urban/industrial soil is anthropogenic. As the impact of industrial development on environment, researches on urban/industrial soil are increasing, and carried out in a wider scale. Researches on particles have become popular because the air is easily impacted by human activities. In some countries, haze has become an important problem. As the main composition, suspended particles in air, especially PM 2.5, are the important media to transport and spread contaminates. The researches on PM 2.5 are carrying out all around the world, both developed and developing countries.

The most popular method used are PCA, FA, and positive matrix fractionation (PMF). The PMF is frequently used in the particle researches, but less in water and soil researches. In the study of soil and particle, semi-supervised ML techniques are also implemented [38]. Some researches combine the machine learning method with geochemical method, or two or more machine learning method together. For example, Petrik et al. [56] combined factor analysis and multivariate linear regression. The ANN is a tool to predict air quality based on history data, relative researches are abundant, Mclean et al. have made a thorough reviewed on this topic [87]. However, very little work has been carried out to identify TE source using ANN method.

From the reviewed reports, anthropogenic source of trace elements in soil and particle includes mainly metal element, Zn, Mn, Ni, Cu, and some other toxic elements, As, Cd, Cr, Hg, Pb, etc. The soil and particle have similar TE composite. More metal TEs are found in soil and particle than that in ground water.

4. Conclusions

The techniques of data mining are widely used to trace sources of TEs in water and solid matrix.

In water environment, ground water and surface water have relation in the flow network. Human activities, especially for the mining, change the natural reaction environment, releasing trace element into ground water and surface water. Then the sediment in river and lake may be contaminated and be a source to water that may release trace element again. Soil, dust, and air particles may be influenced by varies of human activities, especially in the urban and industrial area. The TE composition is different depending on the environmental media type, human activities, land use type, etc. However, some environmental concern element, As, Pb, Cd, Hg, Cr, are frequently found in water, sediment, soil, and particle, showing high mobility and contaminating potential on environment.

The unsupervised machine learning algorithm, including principal component analysis, factor analysis, positive matrix fractionation is mostly used. The PCA is used in water is to find contamination source of trace element, and sometimes water inrush in coal mines. In the air particle researches, PCA and PMF are frequently used to trace the source of PM 2.5 and PM 10, and the TEs source in the particle sources. Some supervised algorithm, including discrimination analysis, Bayesian network, artificial neural network, decision tree is used when the data are labeled.

Generally speaking, the most popular methods used to apportion the source of trace elements as contaminants are unsupervised ML techniques, especially the principal component analysis. In a wider scope, supervised ML is a big tool box for investigations and researches, which is frequently applied and implemented in the areas of science and society. The supervised ML usually gives more accuracy and robust result than the unsupervised ML. In the area of trace element apportionment, some factors constrain the implementation of supervised ML techniques, as the sources are usually not known. However, some techniques are promising to treat the issues of trace element apportionment. First, the supervised ML methods could be implemented more frequently. The unsupervised ML methods are used in the first step. With the intensive research, as some sources have been identified, the supervised ML methods could be used. For example, water inrush is sometimes a threaten in some Chinese coal mines. As the potential source of inrush can be identified, supervised ML method, discriminant analysis is used to determine the water type of inrush, then the corresponding technologies to deal with the threaten or accidents could be implemented. At this stage, some other supervised ML method could also be used. However, the discriminant analysis was mostly used. Second, semi supervised ML may be used implemented more. This method is a series of relative novel techniques. Once more data is obtained in an investigation or research, the semi-supervised ML may be used. In a sense, this method combines the unsupervised and supervised techniques in one implementation. Third, the machine learning method could be combined with geochemical method together. Two technique system have their advantages and disadvantages, the combination may achieve its maximum consequences and efficiency.

Acknowledgements

Our study is funded by the Fundamental Research Funds for the Central Universities (3142014005), the Colleges and Universities in Hebei Province Science and Technology Research Project (ZD2016204), and the National Natural Science Foundation of China (51674119).

Conflict of interest

The authors declare no conflict of interest.

Author details

Yao Shan* and Jianjun Shi
School of Safety Engineering, North China Institute of Science and Technology, Yanjiao, China

*Address all correspondence to: 9106350@qq.com

References

[1] Iqbal J, Shah MH. Distribution, correlation and risk assessment of selected metals in urban soils from Islamabad, Pakistan. Journal of Hazardous Materials. 2011;**192**(2): 887-898. DOI: 10.1016/j.jhazmat.2011.05.105

[2] Manta DS, Angelone M, Bellanca A, Neri R, Sprovieri M. Heavy metals in urban soils: A case study from the city of Palermo (Sicily), Italy. Science of the Total Environment. 2002;**300**(1–3): 229-243

[3] Qu MK, Li WD, Zhang CR, Wang SQ, Yang Y, He LY. Source apportionment of heavy metals in soils using multivariate statistics and geostatistics. Pedosphere. 2013;**23**(4): 437-444. DOI: 10.1016/S1002-0160(13) 60036-3

[4] Xu X, Zhao Y, Zhao X, Wang Y, Deng W. Sources of heavy metal pollution in agricultural soils of a rapidly industrializing area in the Yangtze Delta of China. Ecotoxicology and Environmental Safety. 2014;**108**: 161-167. DOI: 10.1016/j.ecoenv.2014.07.001

[5] Dai S, Li W, Tang Y, Zhang Y, Feng P. The sources, pathway, and preventive measures for fluorosis in Zhijin County, Guizhou, China. Applied Geochemistry. 2007;**22**(5):1017-1024

[6] Hosono T, Su CC, Okamura K, Taniguchi M. Historical record of heavy metal pollution deduced by lead isotope ratios in core sediments from the Osaka Bay, Japan. Journal of Geochemical Exploration. 2010;**107**(1):1-8. DOI: 10.1016/j.gexplo.2010.05.003

[7] Ding F, He Z, Liu S, Zhang S, Zhao F, Li Q, et al. Heavy metals in composts of China: Historical changes, regional variation, and potential impact on soil quality. Environmental Science and Pollution Research. 2017;**24**(3): 3194-3209

[8] Tedoldi D, Chebbo G, Pierlot D, Branchu P, Kovacs Y, Gromaire MC. Spatial distribution of heavy metals in the surface soil of source-control stormwater infiltration devices—Inter-site comparison. Science of the Total Environment. 2017;**579**:881-892. DOI: 10.1016/j.scitotenv.2016.10.226

[9] Sridhara Chary N, Kamala CT, Samuel Suman Raj D. Assessing risk of heavy metals from consuming food grown on sewage irrigated soils and food chain transfer. Ecotoxicology and Environmental Safety. 2008;**69**(3): 513-524

[10] Khan S, Rehman S, Zeb Khan A, Amjad Khan M, Tahir Shah M. Soil and vegetables enrichment with heavy metals from geological sources in Gilgit, northern Pakistan. Ecotoxicology and Environmental Safety. 2010;**73**(7): 1820-1827. DOI: 10.1016/j.ecoenv.2010.08.016

[11] Chabukdhara M, Nema AK. Heavy metals assessment in urban soil around industrial clusters in Ghaziabad, India: Probabilistic health risk approach. Ecotoxicology and Environmental Safety. 2013;**87**:57-64. DOI: 10.1016/j.ecoenv.2012.08.032

[12] Ogunlaja A, Ogunlaja OO, Okewole DM, Morenikeji OA. Risk assessment and source identification of heavy metal contamination by multivariate and hazard index analyses of a pipeline vandalised area in Lagos State, Nigeria. Science of the Total Environment. 2019;**651**:2943-2952. DOI: 10.1016/j.scitotenv.2018.09.386

[13] Yue M, Zhao F. Leaching experiments to study the release of trace elements from mineral separates from

Chinese coals. International Journal of Coal Geology. 2008;**73**(1):43-51

[14] Shan Y, Qin Y, Wang W. Chromium leaching mechanism of coal mine water —A modeling study based on Xuzhou-Datun coal mine district. Mining Science and Technology. 2010;**20**(1):97-102. DOI: 10.1016/S1674-5264(09)60168-X

[15] Khuzestani RB, Souri B. Evaluation of heavy metal contamination hazards in nuisance dust particles, in kurdistan province, western Iran. Journal of Environmental Sciences. 2013;**25**(7): 1346-1354. DOI: 10.1016/S1001-0742 (12)60147-8

[16] Liu J, Yang T, Chen Q, Liu F, Wang B. Distribution and potential ecological risk of heavy metals in the typical eco-units of Haihe River Basin. Frontiers of Environmental Science & Engineering. 2016;**10**(1):103-113. DOI: 10.1007/s11783-014-0686-5

[17] Ma L, Yang Z, Li L, Wang L. Source identification and risk assessment of heavy metal contaminations in urban soils of Changsha, a mine-impacted city in Southern China. Environmental Science and Pollution Research. 2016; **23**(17):17058-17066

[18] Chen T, Liu X, Zhu M, Zhao K, Wu J, Xu J, et al. Identification of trace element sources and associated risk assessment in vegetable soils of the urban-rural transitional area of Hangzhou, China. Environmental Pollution. 2008;**151**(1):67-78

[19] Xia X, Chen X, Liu R, Liu H. Heavy metals in urban soils with various types of land use in Beijing, China. Journal of Hazardous Materials. 2011;**186**(2–3): 2043-2050. DOI: 10.1016/j.jhazmat.2010.12.104

[20] Huang J, Peng S, Mao X, Li F, Guo S, Shi L, et al. Source apportionment and spatial and quantitative ecological risk assessment of heavy metals in soils from a typical Chinese agricultural county. Process Safety and Environment Protection. 2019;**126**:339-347. DOI: 10.1016/j.psep.2019.04.023

[21] Mazurek R, Kowalska J, Gąsiorek M, Zadrożny P, Józefowska A, Zaleski T, et al. Assessment of heavy metals contamination in surface layers of Roztocze National Park forest soils (SE Poland) by indices of pollution. Chemosphere. 2017;**168**:839-850

[22] Chuncai Z, Guijian L, Ting F, Ruoyu S, Dun W. Leaching characteristic and environmental implication of rejection rocks from Huainan Coalfield, Anhui Province, China. Journal of Geochemical Exploration. 2014;**143**:54-61. DOI: 10.1016/j.gexplo.2014.03.010

[23] Li S, Jia Z. Heavy metals in soils from a representative rapidly developing megacity (SW China): Levels, source identification and apportionment. Catena. 2018;**163** (December):414-423. DOI: 10.1016/j.catena.2017.12.035

[24] Anteneh Y, Zeleke G, Gebremariam E. Assessment of surface water quality in legedadie and dire catchments, Central Ethiopia, using multivariate statistical analysis. Acta Ecologica Sinica. 2018;**38**(2):81-95 Available from: https://doi.org/10.1016/j.chnaes.2017.05.005

[25] Varol M. Arsenic and trace metals in a large reservoir: Seasonal and spatial variations, source identification and risk assessment for both residential and recreational users. Chemosphere. 2019; **228**:1-8. DOI: 10.1016/j.chemosphere.2019.04.126

[26] Xue D, De Baets B, Van Cleemput O, Hennessy C, Berglund M, Boeckx P. Use of a Bayesian isotope mixing model to estimate proportional contributions of multiple nitrate sources

in surface water. Environmental Pollution. 2012;**161**:43-49. DOI: 10.1016/j.envpol.2011.09.033

[27] Hajigholizadeh M, Melesse AM. Assortment and spatiotemporal analysis of surface water quality using cluster and discriminant analyses. Catena. 2017; **151**:247-258. DOI: 10.1016/j.catena.2016.12.018

[28] Matiatos I, Paraskevopoulou V, Lazogiannis K, Botsou F, Dassenakis M, Ghionis G, et al. Surface–ground water interactions and hydrogeochemical evolution in a fluvio-deltaic setting: The case study of the Pinios River delta. Journal of Hydrology. 2018;**561**(April):236-249

[29] Matiatos I. Nitrate source identification in groundwater of multiple land-use areas by combining isotopes and multivariate statistical analysis: A case study of Asopos basin (Central Greece). Science of the Total Environment. 2016;**541**:802-814. DOI: 10.1016/j.scitotenv.2015.09.134

[30] Matiatos I, Alexopoulos A, Godelitsas A. Multivariate statistical analysis of the hydrogeochemical and isotopic composition of the groundwater resources in northeastern Peloponnesus (Greece). Science of the Total Environment. 2014;**476–477**:577-590. DOI: 10.1016/j.scitotenv.2014.01.042

[31] Barberá JA, Andreo B. River-spring connectivity and hydrogeochemical interactions in a shallow fractured rock formation. The case study of Fuensanta river valley (Southern Spain). Journal of Hydrology. 2017;**547**:253-268

[32] Huang P, Yang Z, Wang X, Ding F. Research on Piper-PCA-Bayes-LOOCV discrimination model of water inrush source in mines. Arabian Journal of Geosciences. 2019;**12**:334. DOI: 10.1007/s12517-019-4500-3

[33] Wang J, Li X, Cui T, Yang J. Application of distance discriminant analysis method to headstream recognition of water-bursting source. Procedia Engineering. 2011;**26**:374-381. DOI: 10.1016/j.proeng.2011.11.2181

[34] Shan Y, Wang W, Qin Y, Gao L. Multivariate analysis of trace elements leaching from coal and host rock. Groundwater for Sustainable Development. 2019;**8**(November): 402-412. DOI: 10.1016/j.gsd.2019.01.001

[35] Magesh NS, Chandrasekar N, Elango L. Chemosphere trace element concentrations in the groundwater of the Tamiraparani river basin, South India: Insights from human health risk and multivariate statistical techniques. Chemosphere. 2017;**185**:468-479. DOI: 10.1016/j.chemosphere.2017.07.044

[36] Kumar M, Ramanatahn AL, Tripathi R, Farswan S, Kumar D, Bhattacharya P. A study of trace element contamination using multivariate statistical techniques and health risk assessment in groundwater of Chhaprola Industrial Area, Gautam Buddha Nagar, Uttar Pradesh, India. Chemosphere. 2017;**166**:135-145. DOI: 10.1016/j.chemosphere.2016.09.086

[37] Vesselinov VV, Alexandrov BS, Malley DO. Contaminant source identification using semi-supervised machine learning. Journal of Contaminant Hydrology. 2018;**212**(November): 134-142. DOI: 10.1016/j.jconhyd.2017.11.002

[38] Xue D, Pang F, Meng F, Wang Z, Wu W. Decision-tree-model identification of nitrate pollution activities in groundwater: A combination of a dual isotope approach and chemical ions. Journal of Contaminant Hydrology. 2015;**180**: 25-33. DOI: 10.1016/j.jconhyd.2015.07.003

[39] Chen H, Chen R, Teng Y, Wu J. Contamination characteristics,

ecological risk and source identification of trace metals in sediments of the Le'an River (China). Ecotoxicology and Environmental Safety. 2016;**125**:85-92. DOI: 10.1016/j.ecoenv.2015.11.042

[40] Ke X, Gui S, Huang H, Zhang H, Wang C, Guo W. Ecological risk assessment and source identification for heavy metals in surface sediment from the Liaohe River protected area, China. Chemosphere. 2017;**175**:473-481. DOI: 10.1016/j.chemosphere.2017.02.029

[41] Zhang Z, Juying L, Mamat Z. Sources identification and pollution evaluation of heavy metals in the surface sediments of Bortala River, Northwest China. Ecotoxicology and Environmental Safety. 2016;**126**:94-101. DOI: 10.1016/j.ecoenv.2015.12.025

[42] Lin Q, Liu E, Zhang E, Li K, Shen J. Spatial distribution, contamination and ecological risk assessment of heavy metals in surface sediments of Erhai Lake, a large eutrophic plateau lake in Southwest China. Catena. 2016;**145**: 193-203. DOI: 10.1016/j.catena.2016.06.003

[43] Dai L, Wang L, Li L, Liang T, Zhang Y, Ma C, et al. Multivariate geostatistical analysis and source identification of heavy metals in the sediment of Poyang Lake in China. Science of the Total Environment. 2018; **621**:1433-1444. DOI: 10.1016/j.scitotenv.2017.10.085

[44] Qin F, Ji H, Li Q, Guo X, Tang L, Feng J. Evaluation of trace elements and identification of pollution sources in particle size fractions of soil from iron ore areas along the Chao River. Journal of Geochemical Exploration. 2014;**138**: 33-49. DOI: 10.1016/j.gexplo.2013.12.005

[45] Lin Y, Han P, Huang Y, Yuan GL, Guo JX, Li J. Source identification of potentially hazardous elements and their relationships with soil properties in agricultural soil of the Pinggu district of Beijing, China: Multivariate statistical analysis and redundancy analysis. Journal of Geochemical Exploration. 2017;**173**:110-118. DOI: 10.1016/j.gexplo.2016.12.006

[46] Chai Y, Guo J, Chai S, Cai J, Xue L, Zhang Q. Source identification of eight heavy metals in grassland soils by multivariate analysis from the Baicheng-Songyuan area, Jilin Province, Northeast China. Chemosphere. 2015;**134**:67-75. DOI: 10.1016/j.chemosphere.2015.04.008

[47] Silva LFO, Fdez-Ortiz de Vallejuelo S, Martinez-Arkarazo I, Castro K, Oliveira MLS, Sampaio CH, et al. Study of environmental pollution and mineralogical characterization of sediment rivers from Brazilian coal mining acid drainage. Science of the Total Environment. 2013;**447**:169-178. DOI: 10.1016/j.scitotenv.2012.12.013

[48] Lu X, Pan H, Wang Y. Pollution evaluation and source analysis of heavy metal in roadway dust from a resource-typed industrial city in Northwest China. Atmospheric Pollution Research. 2017;**8**(3):587-595. DOI: 10.1016/j.apr.2016.12.019

[49] Shangguan Y, Cheng B, Zhao L, Hou H, Ma J, Sun Z, et al. Distribution assessment and source identification using multivariate statistical analyses and artificial neutral networks for trace elements in agricultural soils in Xinzhou of Shanxi Province, China. Pedosphere. 2018;**28**(3):542-554. DOI: 10.1016/S1002-0160(17)60304-7

[50] Chandrasekaran A, Ravisankar R, Harikrishnan N, Satapathy KK, Prasad MVR, Kanagasabapathy KV. Multivariate statistical analysis of heavy metal concentration in soils of Yelagiri Hills, Tamilnadu, India—Spectroscopical approach. Spectrochimica Acta, Part A: Molecular and Biomolecular Spectroscopy. 2015;

137:589-600. DOI: 10.1016/j.saa.2014.08.093

[51] Tepanosyan G, Sahakyan L, Belyaeva O, Saghatelyan A. Origin identification and potential ecological risk assessment of potentially toxic inorganic elements in the topsoil of the city of Yerevan, Armenia. Journal of Geochemical Exploration. 2016;**167**:1-11. DOI: 10.1016/j.gexplo.2016.04.006

[52] Huang S, Tu J, Liu H, Hua M, Liao Q, Feng J, et al. Multivariate analysis of trace element concentrations in atmospheric deposition in the Yangtze River Delta, East China. Atmospheric Environment. 2009; **43**(36):5781-5790. DOI: 10.1016/j.atmosenv.2009.07.055

[53] Boente C, Matanzas N, García-González N, Rodríguez-Valdés E, Gallego JR. Trace elements of concern affecting urban agriculture in industrialized areas: A multivariate approach. Chemosphere. 2017;**183**: 546-556

[54] Malik RN, Jadoon WA, Husain SZ. Metal contamination of surface soils of industrial city Sialkot, Pakistan: A multivariate and GIS approach. Environmental Geochemistry and Health. 2010;**32**(3):179-191

[55] Kelepertzis E. Accumulation of heavy metals in agricultural soils of Mediterranean: Insights from Argolida basin, Peloponnese, Greece. Geoderma. 2014;**221–222**:82-90. DOI: 10.1016/j.geoderma.2014.01.007

[56] Petrik A, Thiombane M, Albanese S, Lima A, De Vivo B. Source patterns of Zn, Pb, Cr and Ni potentially toxic elements (PTEs) through a compositional discrimination analysis: A case study on the Campanian topsoil data. Geoderma. 2018;**331**(December): 87-99. DOI: 10.1016/j.geoderma.2018.06.019

[57] Liu Y, Ma Z, Lv J, Bi J. Identifying sources and hazardous risks of heavy metals in topsoils of rapidly urbanizing East China. Journal of Geographical Sciences. 2016;**26**(6):735-749

[58] Fatehi M, Asadi HH. Application of semi-supervised fuzzy c-means method in clustering multivariate geochemical data , a case study from the Dalli Cu-Au porphyry deposit in Central Iran. Ore Geology Reviews. 2017;**81**:245-255. DOI: 10.1016/j.oregeorev.2016.10.002

[59] Reeves MK, Perdue M, Munk LA, Hagedorn B. Predicting risk of trace element pollution from municipal roads using site-specific soil samples and remotely sensed data. Science of the Total Environment. 2018;**630**:578-586. DOI: 10.1016/j.scitotenv.2018.02.171

[60] Pandey B, Agrawal M, Singh S. Assessment of air pollution around coal mining area: Emphasizing on spatial distributions, seasonal variations and heavy metals, using cluster and principal component analysis. Atmospheric Pollution Research. 2013;**5**(1):79-86. DOI: 10.5094/APR.2014.010

[61] Bari MA, Kindzierski WB. Fine particulate matter (PM2.5) in Edmonton, Canada: Source apportionment and potential risk for human health. Environmental Pollution. 2016;**218**:219-229. DOI: 10.1016/j.envpol.2016.06.014

[62] Ying Q, Feng M, Song D, Wu L, Hu J, Zhang H, et al. Improve regional distribution and source apportionment of PM 2.5 trace elements in China using inventory-observation constrained emission factors. Science of the Total Environment. 2018;**624**:355-365

[63] Feng J, Yu H, Su X, Liu S, Li Y, Pan Y, et al. Chemical composition and source apportionment of PM 2.5 during Chinese Spring Festival at Xinxiang, a heavily polluted city in North China:

Fireworks and health risks. Atmospheric Research. 2016;**182**:176-188

[64] Zhai Y, Liu X, Chen H, Xu B, Zhu L, Li C, et al. Source identification and potential ecological risk assessment of heavy metals in PM2.5 from Changsha. Science of the Total Environment. 2014; **493**:109-115. DOI: 10.1016/j. scitotenv.2014.05.106

[65] Brinkman GL, Milford JB, Schauer JJ, Shafer MM, Hannigan MP. Source identification of personal exposure to fine particulate matter using organic tracers. Atmospheric Environment. 2009;**43**(12):1972-1981. DOI: 10.1016/j.atmosenv.2009.01.023

[66] Murillo JH, Roman SR, Rojas Marin JF, Ramos AC, Jimenez SB, Gonzalez BC, et al. Chemical characterization and source apportionment of PM10 and PM2.5 in the metropolitan area of Costa Rica, Central America. Atmospheric Pollution Research. 2013;**4**(2):181-190. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1309104215303871

[67] Owoade KO, Hopke PK, Olise FS, Adewole OO, Ogundele LT, Fawole OG. Source apportionment analyses for fine (PM 2.5) and coarse (PM 2.5–10) mode particulate matter (PM) measured in an urban area in southwestern Nigeria. Atmospheric Pollution Research. 2016; **7**(5):843-857. DOI: 10.1016/j. apr.2016.04.006

[68] Sturtz TM, Adar SD, Gould T, Larson TV. Constrained source apportionment of coarse particulate matter and selected trace elements in three cities from the multi-ethnic study of atherosclerosis. Atmospheric Environment. 2014;**84**:65-77. DOI: 10.1016/j.atmosenv.2013.11.031

[69] Zongjie LI, Fei LIU, Yong S, Lingling S, Qing T, Bing JIA, et al. Chemical characteristics of precipitation and the indicative significance for sand dust events in the northern and southern slopes of Wushaoling Mountain , northwestern China. Journal of Arid Land. 2017;**9**:911-923

[70] Bo Y, Liu C, Zhao Y, Wang L. Chemical and isotopic characteristics and origin of spring waters in the Lanping-Simao Basin, Yunnan, Southwestern China. Chemie der Erde —Geochemistry. 2015;**75**(3):287-300. DOI: 10.1016/j.chemer.2015.04.002

[71] Gammons CH, Brown A, Poulson SR, Henderson TH. Using stable isotopes (S, O) of sulfate to track local contamination of the Madison karst aquifer, Montana, from abandoned coal mine drainage. Applied Geochemistry. 2013;**31**:228-238. DOI: 10.1016/j. apgeochem.2013.01.008

[72] Gammons CH, Duaime TE, Parker SR, Poulson SR, Kennelly P. Geochemistry and stable isotope investigation of acid mine drainage associated with abandoned coal mines in Central Montana, USA. Chemical Geology. 2010;**269**(1–2):100-112. DOI: 10.1016/j.chemgeo.2009.05.026

[73] Rao W, Han G, Tan H, Jiang S. Chemical and Sr isotopic compositions of rainwater on the Ordos Desert Plateau, Northwest China. Environment and Earth Science. 2015;**74**(7):5759-5771

[74] Fan Q, Ma H, Lai Z, Tan H, Li T. Origin and evolution of oilfield brines from Tertiary strata in western Qaidam Basin: Constraints from 87Sr/86Sr, δD, δ18O, δ34S and water chemistry. Chinese Journal of Geochemistry. 2010; **29**(4):446-454

[75] Soler A, Canals A, Goldstein SL, Otero N, Antich N, Spangenberg J. Sulfur and strontium isotope composition of the Llobregat river (Ne Spain): Tracers of natural and anthropogenic chemicals in stream waters. Water, Air, and Soil Pollution. 2001;**136**:207-224

[76] Samborska K, Halas S. 34S and 18O in dissolved sulfate as tracers of hydrogeochemical evolution of the Triassic carbonate aquifer exposed to intense groundwater exploitation (Olkusz-Zawiercie region, southern Poland). Applied Geochemistry. 2010; **25**(9):1397-1414. DOI: 10.1016/j.apgeochem.2010.06.010

[77] Bottrell S, Tellam J, Bartlett R, Hughes A. Isotopic composition of sulfate as a tracer of natural and anthropogenic influences on groundwater geochemistry in an urban sandstone aquifer, Birmingham, UK. Applied Geochemistry. 2008;**23**(8): 2382-2394

[78] Cui J, Zang S, Zhai D, Wu B. Potential ecological risk of heavy metals and metalloid in the sediments of Wuyuer River basin, Heilongjiang Province, China. Ecotoxicology. 2014; **23**(4):589-600

[79] Maanan M, Saddik M, Maanan M, Chaibi M, Assobhei O, Zourarah B. Environmental and ecological risk assessment of heavy metals in sediments of Nador lagoon, Morocco. Ecological Indicators. 2015;**48**:616-626. DOI: 10.1016/j.ecolind.2014.09.034

[80] Ujević Bošnjak M, Capak K, Jazbec A, Casiot C, Sipos L, Poljak V, et al. Hydrochemical characterization of arsenic contaminated alluvial aquifers in Eastern Croatia using multivariate statistical techniques and arsenic risk assessment. Science of the Total Environment. 2012;**420**:100-110

[81] Taner S, Pekey B, Pekey H. Fine particulate matter in the indoor air of barbeque restaurants: Elemental compositions, sources and health risks. Science of the Total Environment. 2013; **454–455**:79-87. DOI: 10.1016/j.scitotenv.2013.03.018

[82] Massey DD, Kulshrestha A, Taneja A. Particulate matter concentrations and their related metal toxicity in rural residential environment of semi-arid region of India. Atmospheric Environment. 2013;**67**:278-286. DOI: 10.1016/j.atmosenv.2012.11.002

[83] Cánovas CR, Olías M, Macias F, Torres E, San Miguel EG, Galván L, et al. Water acidification trends in a reservoir of the Iberian Pyrite Belt (SW Spain). Science of the Total Environment. 2016;**541**:400-411. DOI: 10.1016/j.scitotenv.2015.09.070

[84] Erhardt EB, Bedrick EJ. A Bayesian framework for stable isotope mixing models. Environmental and Ecological Statistics. 2013;**20**(3):377-397

[85] Qian SS, Craig JK, Baustian MM, Rabalais NN. A Bayesian hierarchical modeling approach for analyzing observational data from marine ecological studies. Marine Pollution Bulletin. 2009;**58**(12):1916-1921. DOI: 10.1016/j.marpolbul.2009.09.029

[86] Qian SS, Miltner RJ. A continuous variable Bayesian networks model for water quality modeling: A case study of setting nitrogen criterion for small rivers and streams in Ohio, USA. Environmental Modelling & Software. 2015;**69**:14-22. DOI: 10.1016/j.envsoft.2015.03.001

[87] Mclean S, Kaiser J, Richard B. Environmental modelling & software a review of artificial neural network models for ambient air pollution prediction. Environmental Modelling & Software, June. 2019;**119**:285-304. DOI: 10.1016/j.envsoft.2019.06.014

[88] Kilic E, Yucel N. Determination of spatial and temporal changes in water quality at Asi River using multivariate statistical techniques. Turkish Journal of Fisheries and Aquatic Sciences. 2018; **19**(9):727-737

[89] Seal RR, Shanks WC. Sulfide oxidation: Insights from experimental,

theoretical, stable isotope, and predictive studies in the field and laboratory. Applied Geochemistry. 2008;**23**(2):101-102. DOI: 10.1016/j.apgeochem.2007.10.006

[90] Liu P, Hoth N, Drebenstedt C, Sun Y, Xu Z. Hydro-geochemical paths of multi-layer groundwater system in coal mining regions—Using multivariate statistics and geochemical modeling approaches. Science of the Total Environment. 2017;**601–602**:1-14. DOI: 10.1016/j.scitotenv.2017.05.146

[91] Liu B, Tang Z, Dong S, Wang L, Liu D. Vegetation recovery and groundwater pollution control of coal gangue field in a semi-arid area for a field application. International Biodeterioration and Biodegradation. 2018;**128**:134-140. DOI: 10.1016/j.ibiod.2017.01.032

[92] Sahoo PK, Tripathy S, Panigrahi MK, Equeenuddin SM. Geochemical characterization of coal and waste rocks from a high sulfur bearing coalfield, India: Implication for acid and metal generation. Journal of Geochemical Exploration. 2014;**145**: 135-147. DOI: 10.1016/j.gexplo.2014.05.024

[93] Park JH, Edraki M, Baumgartl T. A practical testing approach to predict the geochemical hazards of in-pit coal mine tailings and rejects. Catena. 2017;**148**: 3-10. DOI: 10.1016/j.catena.2015.10.027

[94] Hendryx M. The extractive industries and society the public health impacts of surface coal mining. Extractive Industries and Society. 2015;**2**(4):820-826. DOI: 10.1016/j.exis.2015.08.006

[95] Cravotta CA. Monitoring, field experiments, and geochemical modeling of Fe(II) oxidation kinetics in a stream dominated by net-alkaline coal-mine drainage, Pennsylvania, USA. Applied Geochemistry. 2015;**62**:96-107. DOI: 10.1016/j.apgeochem.2015.02.009

[96] Nordstrom DK. Hydrogeochemical processes governing the origin, transport and fate of major and trace elements from mine wastes and mineralized rock to surface waters. Applied Geochemistry. 2011;**26**(11): 1777-1791. DOI: 10.1016/j.apgeochem.2011.06.002

[97] Tozsin G. Hazardous elements in soil and coal from the Oltu coal mine district, Turkey. International Journal of Coal Geology. 2014;**131**:1-6. DOI: 10.1016/j.coal.2014.05.011

[98] Martins-Ferreira MAC, Campos JEG, Pires ACB. Near-mine exploration via soil geochemistry multivariate analysis at the almas gold province, Central Brazil: A study case. Journal of Geochemical Exploration. 2017;**173**:52-63. DOI: 10.1016/j.gexplo.2016.11.011

[99] Oliveira MLS, Ward CR, French D, Hower JC, Querol X, Silva LFO. Mineralogy and leaching characteristics of beneficiated coal products from Santa Catarina, Brazil. International Journal of Coal Geology. 2012;**94**:314-325. DOI: 10.1016/j.coal.2011.10.004

[100] Raja R, Nayak AK, Shukla AK, Rao KS, Gautam P, Lal B, et al. Impairment of soil health due to fly ash-fugitive dust deposition from coal-fired thermal power plants. Environmental Monitoring and Assessment. 2015;**187**: 679. DOI: 10.1007/s10661-015-4902-y

[101] Munawer ME. Human health and environmental impacts of coal combustion and post-combustion wastes. Journal of Sustainable Mining. 2018;**17**(2):87-96. DOI: 10.1016/j.jsm.2017.12.007

[102] Schwartz GE, Rivera N, Lee SW, Harrington JM, Hower JC, Levine KE, et al. Leaching potential and redox transformations of arsenic and selenium in sediment microcosms with fly ash. Applied Geochemistry. 2016;**67**:177-185. DOI: 10.1016/j.apgeochem.2016.02.013

[103] Pandey VC, Singh JS, Singh RP, Singh N, Yunus M. Arsenic hazards in coal fly ash and its fate in Indian scenario. Resources, Conservation and Recycling. 2011;**55**(9–10):819-835. DOI: 10.1016/j.resconrec.2011.04.005

[104] Lokeshappa B, Dikshit AK. Fate of metals in coal fly ash ponds. International Journal of Environmental Science and Development. 2013;**1** (January):43-48. DOI: 10.1016/j. apcbee.2012.03.007

[105] Pumure I, Renton JJ, Smart RB. The interstitial location of selenium and arsenic in rocks associated with coal mining using ultrasound extractions and principal component analysis (PCA). Journal of Hazardous Materials. 2011; **198**:151-158. DOI: 10.1016/j. jhazmat.2011.10.032

[106] Yi Y, Yang Z, Zhang S. Ecological risk assessment of heavy metals in sediment and human health risk assessment of heavy metals in fishes in the middle and lower reaches of the Yangtze River basin. Environmental Pollution. 2011;**159**(10):2575-2585. DOI: 10.1016/j.envpol.2011.06.011

[107] Tian HZ, Zhu CY, Gao JJ, Cheng K, Hao JM, Wang K, et al. Quantitative assessment of atmospheric emissions of toxic heavy metals from anthropogenic sources in China: Historical trend, spatial distribution, uncertainties, and control policies. Atmospheric Chemistry and Physics. 2015;**15**(17):10127-10147

[108] Lu Y, Zhu F, Chen J, Gan H, Guo Y. Chemical fractionation of heavy metals in urban soils of Guangzhou, China. Environmental Monitoring and Assessment. 2007;**134**(1–3):429-439

[109] Sheng J, Wang X, Gong P, Tian L, Yao T. Heavy metals of the Tibetan top soils: Level, source, spatial distribution, temporal variation and risk assessment. Environmental Science and Pollution Research. 2012;**19**(8):3362-3370

[110] Skousen JG, Ziemkiewicz PF, McDonald LM. Acid mine drainage formation, control and treatment: Approaches and strategies. Extractive Industries and Society. 2019;**6**(1):241-249

[111] Solgi E, Abbas ES, Alireza RB, Hadipour M. Soil contamination of metals in the three industrial estates, Arak, Iran. Bulletin of Environmental Contamination and Toxicology. 2012; **88**(4):634-638

[112] Hovmand MF, Kemp K, Kystol J, Johnsen I, Riis-Nielsen T, Pacyna JM. Atmospheric heavy metal deposition accumulated in rural forest soils of southern Scandinavia. Environmental Pollution. 2008;**155**(3):537-541

[113] Nziguheba G, Smolders E. Inputs of trace elements in agricultural soils via phosphate fertilizers in European countries. Science of the Total Environment. 2008;**390**(1):53-57

[114] Hou D, O'Connor D, Nathanail P, Tian L, Ma Y. Integrated GIS and multivariate statistical analysis for regional scale assessment of heavy metal soil contamination: A critical review. Environmental Pollution. 2017;**231**: 1188-1200. DOI: 10.1016/j. envpol.2017.07.021

[115] Acosta JA, Faz A, Martínez-Martínez S, Arocena JM. Enrichment of metals in soils subjected to different land uses in a typical Mediterranean environment (Murcia City, Southeast Spain). Applied Geochemistry. 2011; **26**(3):405-414. DOI: 10.1016/j. apgeochem.2011.01.023

[116] Atafar Z, Mesdaghinia A, Nouri J, Homaee M, Yunesian M, Ahmadimoghaddam M, et al. Effect of fertilizer application on soil heavy metal concentration. Environmental Monitoring and Assessment. 2010;**160**(1–4):83-89

[117] Fu Z, Zhai Y, Wang L, Zeng G, Li C, Peng W, et al. Morphological,

geochemical composition and origins of near-surface atmospheric dust in Changsha city of China. Environment and Earth Science. 2012;**66**(8): 2207-2216

[118] Liacos JW, Kam W, Delfino RJ, Schauer JJ, Sioutas C. Characterization of organic, metal and trace element PM2.5 species and derivation of freeway-based emission rates in Los Angeles, CA. Science of the Total Environment. 2012;**435–436**:159-166

[119] Vecchi R, Marcazzan G, Valli G. A study on nighttime-daytime PM10 concentration and elemental composition in relation to atmospheric dispersion in the urban area of Milan (Italy). Atmospheric Environment. 2007;**41**(10):2136-2144

[120] Taneepanichskul N, Gelaye B, Grigsby-Toussaint DS, Lohsoonthorn V, Jimba M, Williams MA. Short-term effects of particulate matter exposure on daily mortality in Thailand: A case-crossover study. Air Quality, Atmosphere, and Health. 2018;**11**(6): 639-647

[121] Boman J, Wagner A, Gatari MJ. Trace elements in PM2.5 in Gothenburg, Sweden. Spectrochimica Acta Part B: Atomic Spectroscopy. 2010;**65**(6): 478-482. DOI: 10.1016/j.sab.2010.03.014

[122] Zhang J, Zhou X, Wang Z, Yang L, Wang J, Wang W. Trace elements in PM 2.5 in Shandong Province: Source identification and health risk assessment. Science of the Total Environment. 2018;**621**:558-577. DOI: 10.1016/j.scitotenv.2017.11.292

[123] Bellido-Martín A, Gómez-Ariza JL, Smichowsky P, Sánchez-Rodas D. Speciation of antimony in airborne particulate matter using ultrasound probe fast extraction and analysis by HPLC-HG-AFS. Analytica Chimica Acta. 2009;**649**(2):191-195

[124] Belis CA, Karagulian F, Larsen BR, Hopke PK. Critical review and meta-analysis of ambient particulate matter source apportionment using receptor models in Europe. Atmospheric Environment. 2013;**69**:94-108. DOI: 10.1016/j.atmosenv.2012.11.009

[125] Karagulian F, Belis CA, Dora CFC, Prüss-Ustün AM, Bonjour S, Adair-Rohani H, et al. Contributions to cities' ambient particulate matter (PM): A systematic review of local source contributions at global level. Atmospheric Environment. 2015;**120**: 475-483. DOI: 10.1016/j.atmosenv.2015.08.087