Effectiveness of Predicting The Spatial Distributions of Target Contaminants of a Coking Plant Based On Their Related Pollutants

Pengwei Qiao
Institute of Resources and Environment, Beijing Academy of Science and Technology, Beijing Key Laboratory of Remediation of Industrial Pollution Sites

Donglin Lai
YuHuan Environmental Technology Co., Ltd

Sucai Yang (1598073927@qq.com)
Institute of Resources and Environment, Beijing Academy of Science and Technology, Beijing Key Laboratory of Remediation of Industrial Pollution Sites

Qianyun Zhao
YuHuan Environmental Technology Co., Ltd

Hengqin Wang
YuHuan Environmental Technology Co., Ltd

Research Article

Keywords: Cokriging, Auxiliary pollutants, Prediction accuracy, Effectiveness, Coking plant, Heavy metal, Polycyclic aromatic hydrocarbons

Posted Date: September 21st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-880779/v1

Version of Record: A version of this preprint was published at Environmental Science and Pollution Research on January 15th, 2022. See the published version at https://doi.org/10.1007/s11356-021-17951-z.

Abstract

The prediction accuracy of the spatial distribution of soil pollutants at a site is relatively low. Related pollutants can be used as auxiliary variables to improve the prediction accuracy. However, little relevant research has been conducted on site soil pollution. To analyze the prediction accuracy of target pollutants combined with auxiliary pollutants, Cu, toluene, and phenanthrene were selected as the target pollutants for this study. Based on geostatistical analysis and spatial analysis, the following results were obtained. (1) The reduction rate of the root mean square errors (RMSEs) for Cu, toluene, and phenanthrene with multivariable cokriging were 68.4%, 81.6%, and 81.2%, respectively, which are proportional to the correlation coefficient of the relationship between the auxiliary pollutants and the target pollutants. (2) The predicted results for Cu, phenanthrene, and toluene and their corresponding related pollutants are more accurate than the results obtained not using the related pollutants. (3) In the interpolation process, the RMSEs for Cu, toluene, and phenanthrene with multivariable cokriging basically increase as the neighborhood sample data increases, and then they become stable. (4) When 84, 61, and 34 sample points were removed, the RMSEs for Cu, toluene, and phenanthrene, respectively with multivariable cokriging were close to the RMSEs of the target pollutants based on the total samples. The results are of great significance to improving the prediction accuracy of the spatial distribution of soil pollutants at coking plant sites.

Introduction

Contaminated sites have received worldwide attention, and China is facing serious problems from polluted sites (Li et al. 2017; Roslund et al. 2018; Shi et al. 2017; Cui et al. 2017). According to the national soil pollution survey bulletin of China, the soil environmental problems in industrial and mining wastelands are prominent. The over-standard rate of the soil pollutant content in heavily polluting enterprises, industrial wastelands, industrial parks, centralized solid waste disposal sites, oil production areas, and mining areas is as high as 20–40%. Contaminated soil has significant adverse effects on human health and the environment(Gou et al. 2019; Yang et al. 2018). Therefore, remediation and treatment are needed for those sites (Liu et al. 2015; Fang et al. 2017).

Accurate knowledge of the spatial distribution of pollutants is key to delineating the scope of remediation efforts (Liu et al. 2017). Overestimating the scope of the remediation increases the cost of the remediation, while underestimating the scope fails to eliminate the risks to human health caused by contaminated sites in an comprehensive way (Li and Heap 2014). The restoration boundary is delimited based on a contour line, which is obtained using spatial interpolation methods and pollutant concentration data from a limited number of soil samples collected throughout the site (Ren et al. 2016; Ma et al. 2016). Commonly used interpolation methods include the ordinary kriging (OK), inverse distance weighted (IDW), and radial basis function (RBF) methods (Dong et al. 2011; Chen et al. 2016; Wu et al. 2008; Gutierrez et al. 2015; Wu et al. 2013; Goovaerts et al. 2008; Carlon et al. 2001; Saito and Goovaerts 2000; Qiao et al. 2018). Their prediction accuracy was has been relatively low for sites compared to that for large scale farmland (Santos-Francés et al. 2017; Weindorf et al. 2013; Paulette et al. 2015). This is

due to the strong spatial variability of soil pollutants at sites, which causes uncertainty (Saito and Goovaerts 2002; Hofmann et al. 2010).

The strong spatial variability and localization of the characteristics of the soil pollution of a site often occur in response to the production and pipeline layout, the pollution source distribution, the pollutant properties, and the soil properties (Qiao et al. 2019; Wu et al. 2011; Liu et al. 2013). There are three main types of soil pollution, each with local site characteristics that are noticeable, such as (Armiento et al. 2011; Girault et al. 2016; Monaco et al. 2015): 1) a small number of high peak values exist in isolation; 2) the concentration of the pollutants decreases sharply during the transition from extremely high values to low values in a neighborhood; and 3) the change in the pollutant concentration in other regions is relatively small. The smoothing effect and the global single spatial gradient expression of the commonly used interpolation models blur the local features and reduce the accuracy of the prediction, which is not conducive to the delimitation of the remediation boundary (Xie et al. 2011; Huo et al. 2010; Robinson and Metternicht 2006; Ding et al. 2017).

In order to improve the prediction accuracy, normal transformation was applied to the highly skewed data for a site (Juang et al. 2001; Wu et al. 2006), multiple interpolation models were jointly used in the different subregions (Wu et al. 2011), and auxiliary factors (e.g., topographic features, pH values, and organic matter contents) were combined using the kriging method (Fu et al. 2018; Vyas et al. 2004; Schnabel and Tietje 2003). However, the presence of several high peak values caused the difficulties in the normal transformation (Wu et al. 2011), and the limited sampling points decreased the accuracy of the interpolation in the subregions (Modis et al. 2008). A cokriging method with auxiliary variables can be used to obtain relatively accurate spatial distribution information for pollutants (Le et al. 2019; Tziachris et al. 2017). This is because cokriging method can incorporate both spatial and intervariable correlations into the spatial interpolation (Juang et al. 1996). Auxiliary factors can be used to calibrate the prediction results of the target pollutants to a certain extent. Nevertheless, the acquisition of auxiliary variables increases the sampling and analysis workload.

The pollutants related to the target pollutant are an easily achieved auxiliary factor. According to the sampling and measurement processes, after the soil samples are obtained, the procedures for analyzing various pollutants of the same type, e.g., heavy metals pollutants (HMs), semi-volatile organic pollutants (SVOCs), and volatile organic pollutants (VOCs), are basically the same, and their data can be obtained almost simultaneously (USEPA 2014, 2018, 2017). Moreover, the sources of multiple pollutants at the same site are roughly the same, which leads to correlations among the various pollutants (Li et al. 2013; Rashed 2010). Therefore, the related pollutants have the potential to be a good source of auxiliary data for use in the cokriging method (Yang et al. 2016). However, it is unclear how much influence cokriging using relevant pollutants has on the accuracy of the prediction since little relevant research has been done to analyze this problem in detail (Zhang and Yang 2017; Juang and Lee 1998).

In this study, we take three types of pollutants (HMs, SVOCs, and VOCs) in the soils of an abandoned coking plant in the central part of Hebei Province, China, as the target pollutants, and the prediction

effects of cokriging using related pollutants of the same type were analyzed. This study was conducted (1) to evaluate the extent of the accuracy improvement in the accuracy of the cokriging analysis by using related pollutants compared to the ordinary kriging; and (2) to analyze the uncertainty in the results of the application of cokriging using related pollutants. The results of this study provide a useful method for determining the spatial distribution prediction of pollutants in industrial sites.

Materials And Methods

2.1 Sample collection and analysis

This study focuses on an abandoned coking plant with 0.175 km^2 in Hebei province, China, which mainly produced coke, coal gas, coal tar, ammonia and some other coal chemical products. The abandoned manufactured gas plant began operating started in 1988 and closed in 2017. In 2018, an environmental investigation of the site was conducted, and 300 soil samples were collected from 126 sample points using direct-push drilling methods (Fig. 1).

For each sample, about 5 g of soil was collected in a nonperturbed way and was placed in a 40-mL headspace bottle with 5 mL of methanol in preparation to analyze the VOCs. Approximately 300 g of soil was placed in a 250-mL glass bottle for the heavy metal and semi-volatile organic pollutant analyses. Subsequently, all of the samples were placed in a sample storage box at 4°C, and then, they were immediately sent to the laboratory for monitoring and analysis. The HMs, SVOCs, and VOCs in the soil samples were determined respectively the following United States Environmental Protection Agency Methods: 6020B (ERF) (USEPA 2014), 8270E (ERF) (USEPA 2018), 8260D (ERF) (USEPA, 2017), respectively.

2.2 The cokriging principle

Cokriging is a geostatistical method used to estimate the target variables using measurement data for several auxiliary variables. It not only combines the autocorrelation of the target variable and the auxiliary variable, but it also determines the relationships between multiple auxiliary variables and the target variable. If $Z_1,...,Z_m$ is the value of $Z_1(x),..., Z_m(x)$ at point x, and $Z(x) = [Z_1(x),...,Z_m(x)]$, then the formula of cokriging is

$$Z^*(x) = \sum_{i=1}^{n} Z(x_i)\Gamma_i \quad \text{Eq. 1}$$

where Γ_i is the weight vector. In order to determine Γ_i, it is necessary to consider the variable $Z_j(j=1,2,...,m)$ at position x as a random variable. It is also required that the cokriging estimation formula provides an unbiased estimation, and the sum of the estimation variance of the corresponding variables is the minimum.

The fitting of the cross semi-variogram is used to describe the spatial continuity of the crossover between the different variables. The formula for calculating the cross-semi-variogram is shown as **Eq. 2** (Li et al., 2013; Zhou et al., 2016; Lu et al., 2012).

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [a_i(x_i) - a_i(x_i + h)][b_i(x_i) - b_i(x_i + h)] \quad \textbf{Eq. 2}$$

where h is the separation distance or lag distance, $N(h)$ is the number of data pairs for the separation distance h, a_i and b_i indicate the primary and the target pollutant and the related pollutant, respectively.

The interpolation formula for predicting the spatial distribution of the target pollutant based on the related pollutant estimates the value at location x_0 as shown in **Eq. 3**.

$$Z^*(x_0) = \sum_{i=1}^{N} a_i Z(x_i) + \sum_{j=1}^{N} b_j Z(x_j) \quad \textbf{Eq. 3}$$

where, Z* (x_0) is the predicted value, N is the number of samples, a_i and b_j are the cokriging weights, Z(x_i) is the measured value of the property of the target pollutant and Z(x_j) is the value of the related pollutant, respectively.

Eq. 3 shows that Z* (x_0) is obtained by weighted averaging of the observations of Z(x_i) and Z(x_j) at their spatial locations. The weighting coefficients (a_i and b_j) are determined based on the distance between the measurement points of the two variables and location x_0, the distance between the measurement points, the semi-variograms of the two variables and their cross-semi-variograms. Thus, in the weighted averaging of the two measurements, the auxiliary information provided by the related pollutants can be used to indicate the spatial variation characteristics of the target pollutants.

2.3 The principle for determination of target pollutants and their related analysis

In order to better analyze the prediction effect of the spatial distribution of the target pollutants based on the related pollutants, in this study, we chose the target pollutants based on the following principles.

Abnormal values have a relatively high deviation, which reflects the characteristics of the soil pollution of the industrial site. In this section, the level of deviation was characterized by the ratio of the highest concentration to the sum of the mean plus three times the standard deviation (abbreviated as $R_{h/sum}$) (**Eq. 4**).

$R_h/sum = C_{highest} / (Mean + 3 \times SD)$ **Eq. 4**

where $R_{h/sum}$ is the level of deviation. A high $R_{h/sum}$ value indicates that the abnormal value deviates far from the norm. The *Mean* is the mean values of all of the concentration data; and *SD* is the standard deviation of all of the concentration data. Values of greater than $(\mathrm{Mean} + 3 \times \mathrm{SD})$ are defined as extreme outliers.

The variability in the pollution data is stronger for the respective types, which is characterized by the CV. This is another characteristics of soil pollution at industrial sites. The higher deviation in the abnormal values and a stronger variability in the pollution data reflect the characteristics of the soil pollutants of the industrial site.

There are several related pollutants, which are strongly correlated with the target pollutants. These related pollutants can be used as the auxiliary pollutants in the spatial interpolation of the target pollutants.

2.4 Data analysis method

The correlation analysis was performed using SPSS 16.0. The semi-variogram model was fitted using the GS + 10.0 and 'gstat' package. The spatial interpolation and cross-validation were performed using ArcGIS 10.1. The root mean square error (RMSE) was used as the index of the cross validation (**Eq. 5**). The accuracy of spatial interpolation is higher when the value of the RMSE is lower. The variability is characterized by a coefficient of variance (CV). When CV < 0.1, the spatial variability is weak; and when 0.1 < CV < 0.9, the spatial variability is moderate (Gao and Shao 2012).

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum\left[Z(x_i) - Z^*(x_i)\right]^2} \quad \textbf{Eq. 5}$$

where $Z(x_i)$ is the measured concentration at location x_i; $Z^*(x_i)$ is the predicted concentration based on the interpolation method at location x_i and n is the number of samples.

The cross-validation conducted in this research refers to the "leave one out" method, which involves the following steps. (1) Temporarily remove a measurement value $Z(x_1)$ from the data set. (2) Estimate the value of $Z^*(x_1)$ at point x_1 using the other measurements based on the interpolation model. (3) Add the data describing the pollutant concentration at point x_1 back into the data set. (4) Repeat step (1) through (3) to estimate the concentration at the other points separately and obtain the estimated value of each point ($Z^*(x_2)$, $Z^*(x_3)$, …, $Z^*(x_n)$). (5) Calculate the RMSE (**Eq. 5**) to assess the accuracy of the prediction results (Qiao et al. 2019).

Results

3.1 Descriptive statistics of target pollutants

According to the principles in Sect. 2.3, Cu, phenanthrene, and toluene, which are HMs, SVOC and VOC pollutants, respectively, were selected as the target pollutants. The auxiliary pollutants were determined based on the results of the correlation analysis, which revealed a significant correlation with the target pollutants at the 0.01 level. The auxiliary pollutants for Cu are Cd, As, Pb and Ni, for phenanthrene are fluoranthene, pyrene, chrysene and benzaanthracene, for toluene are benzene, styrene and m/p-xylene. The descriptive statistics of these target pollutants, such as their minimum values (Min), maximum values (Max), CVs and $R_{h/sum}$ values, are shown in Fig. 2.

3.2 Degree of improvement in the prediction accuracy of the spatial distribution trend of the target pollutants in combining with their various related pollutants

The prediction accuracies of the three target pollutants in combination with their various related pollutants are shown as Fig. 3. When combined with the related pollutants, the prediction error which is indicated by RMSE of the spatial distribution of the target pollutants obtained using single variable cokriging is lower. For Cu combined with As, the reduction in the RMSE was approximately 37%. For toluene combined with benzene, the reduction in the RMSE was 76.8%. For phenanthrene combined with four related pollutants, the reduction in the RMSE was approximately 79%.

However, the reduction in the RMSE for Cu combined with Ni and Pb, toluene combined with M/p-xylene, and phenanthrene combined with benzaanthracene were lower. Therefore, Ni and Pb were not used as auxiliary pollutants for Cu in the multivariable cokriging; M/p-xylene was not used for toluene, and benzaanthracene was not used for phenanthrene. The reductions in RMSEs of Cu, toluene, and phenanthrene using multivariable cokriging were 68.4%, 81.6%, and 81.2%, respectively. Moreover, the RMSEs calculated for the multivariable cokriging were lower than those obtained by only combining one related pollutants.

3.3 The prediction accuracies of the high value areas of the three target pollutants combined with their related pollutants

3.2.1 General distribution trend of the interpolation results

The contour lines in Fig. 4 were extracted from the raster layer predicted using the OK method. The region in Fig. 4 with the dense contour distribution is the area with the high concentration data predicted using the OK method. The contour interval is the same in Fig. 4(I–VI) was same. This is also true for Fig. 4(a–f) and Fig. 4(i–v).

Overall, the spatial distributions of Cu, toluene, and phenanthrene predicted using the OK interpolation and using the related pollutants are similar. The correlation coefficients between the raster layers interpolated using the OK method for Cu with and without the related pollutants are all greater than 0.7; while the correlation coefficients of the raster layers for phenanthrene and toluene are all greater than 0.9. Among them, the spatial transition in the Cu content predicted using the related pollutants (Fig. 4(II–VI)) is relatively gentle compared with that predicted without the related pollutants (Fig. 4(I)). Furthermore, the long axis direction of the Cu and phenanthrene distributions are both northwest-southeast, while the higher concentration of toluene are located in the northern part of the study area.

3.2.2 Prediction accuracy of the high value area

In the process of the investigation and remediation of a contaminated site, the recognition accuracy of the high pollution area is the main concern. In this study, the phenanthrene contents of 5.08% of the sample points exceeded the screening level for the soil environmental risk assessment of sites (DB11T811-2011) (40 mg/kg) (DB11/T 811 2011), while none of the Cu and toluene content of the samples sites exceed their corresponding screening values. In order to analyze the differences in the high value region prediction results of the interpolation methods with different combinations of the related pollutants, the 93rd percentile (80 mg/kg) of the Cu content and the 75th percentile (9 mg/kg) of the toluene content, which make the recognition area easy to identify and analyze, were taken as the boundary of the high value region. The boundaries of the high value region are denoted by the labeled red lines in Fig. 4.

According to Fig. 4, the positions exceeding 40 mg/kg, 80 mg/kg, and 9 mg/kg of Cu, toluene, and phenanthrene, respectively, predicted with and without using the related pollutants are the same, but the area is different. The area enclosed by the Cu contour without related pollutants is the smallest, and the area for the combination of Cu and Pb is the largest. In addition, the area for a combination of Cu with Cd and As falls between those for only Cu and for Cu and Pb. For phenanthrene, the area for the combination of fluoranthene, pyrene, and chrysene is the smallest compared with Fig. 4(a–e). For toluene, the area for the combination of toluene and styrene is the smallest, and the area for the combination of toluene and benzene is the largest. In addition, the area for the combination of toluene with benzene and styrene falls between the areas of the other two combinations.

The prediction accuracy of the area of the different combinations, i.e., the prediction accuracy of the high value area, was judged based on the number of sampling points with values greater than the standard pollutant content values inside and outside the high value area. According to the statistical results, of the 118 sampling points, 9 samples had Cu contents greater than 80 mg/kg, 6 samples had phenanthrene contents greater than 40 mg/kg, and 29 samples had toluene contents greater than 9 mg/kg.

According to Table 1, all 9 of the samples with Cu contents of greater than 80 mg/kg were identified using cokriging and a combination of Cu and Pb, and using a combination of Cu and both Cd and As. However, the high pollution area predicted using the combination of Cu and both Cd and As was smaller than that predicted using the combination of Cu and Pb (Fig. 4(III–VI)). Therefore, the locations with Cu

and both Cd and As, and the prediction area of the high value region is small, which lowers the of the remediation. Thus, the prediction results are ideal.

For phenanthrene, although all of the points with phenanthrene contents of greater than 40 mg/kg were accurately identified, the points not exceeding 40 mg/kg were also included in the high value area. The error recognition rate of the prediction results of the combination of phenanthrene with benzaanthracene, chrysene, pyrene, and fluoranthene was approxinately 50%, while that of the combination of phenanthrene and fluoranthene, pyrene, and chrysene was only about 33% (Table 1). In addition, the prediction area of the high value region is the smallest (Fig. 4(a–f)). Therefore, the prediction result for the combination of phenanthrene with fluoranthene, pyrene, and chrysene is more accurate, and the repair cost is the lowest.

The error recognition rates for toluene were all 60%, except for the combination of toluene with benzene and styrene (67%) (Table 1). However, the sample points with toluene contents of greater than 9 mg/kg were distributed in the high value area predicted in the central region as shown in Fig. 4(i–v), which is the region that was predicted accurately. The areas of the regions that were predicted accurately in Fig. 4(i–v) are 23485 m^2, 22488 m^2, 21278 m^2, 23680 m^2, and 20876 m^2, respectively. Therefore, the prediction result of toluene combined with benzene and styrene was better than the prediction result of toluene combined with other related pollutants.

Table 1
Statistical results for the number of sample points with values greater than the standard pollutant content values in the high value area predicted using different combinations of related pollutants

Target pollutant	Interpolation combination	Number of sample points in the high value area	Number of sample points with values greater than the standard pollutant content values in the high value area	The total number of sample points with values greater than the standard pollutant content values for 118 sample points
Cu	Cu	8	8	9
	Cu-Ni	8	8	9
	Cu-Pb	9	9	9
	Cu-As	8	8	9
	Cu-Cd	8	8	9
	Cu-Cd-As	9	9	9
Ph	Ph	12	5	6
	Ph-Benza	12	6	6
	Ph-Ch	12	6	6
	Ph-Py	13	6	6
	Ph-Fl	12	6	6
	Ph-Fl-Py-Ch	9	6	6
To	To	5	2	2
	To_M/P	5	2	2
	To_St	5	2	2
	To-Benze	6	2	2
	To-Benze-St	5	2	2
Ph: phenanthrene; Benza: benzaanthracene; Ch; chrysene; Py: pyrene; Fl: fluoranthene;				

To: toluene; M/p: m/p-xylene; St: styrene; Benze: benzene.

Discussion

4.1 The effect of the correlation between the content of the related pollutants and the contents of the target pollutants

In this study, there is a significant positive correlation between the contents of the related pollutants and the contents of target pollutants, which can be used to calibrate the descriptive results of the spatial distribution characteristics of the target pollutants to some extent. The reduction in the RMSE of the spatial distribution of the target pollutants combined with the related pollutants increases as the correlation between the related pollutants and the target pollutants increases (Fig. 5). This result is consistent with the results of previous studies (Khosravi and Balyani 2019).

4.2 The influence of the number of neighboring samples on the prediction accuracy

Based on the predicted results, for combinations of auxiliary pollutants, the prediction error of the target pollutants is reduced, i.e., RMSE of Cu > RMSE of Cu-auxiliary pollutants, RMSE of toluene > RMSE of toluene-auxiliary pollutants, RMSE of phenanthrene > RMSE of phenanthrene-auxiliary pollutants. The advantages of combining the target pollutant with auxiliary pollutants is demonstrated by the results of this study (Zhen et al. 2019; Wu et al. 2006).

In addition, the prediction error of a target pollutant combined with multiple correlation auxiliary pollutants is smaller than that of combination with a single correlation auxiliary pollutant, i.e., RMSE of Cu-Cd-As < RMSE of Cu-Cd and RMSE of Cu-As, RMSE of toluene-benzene-styrene < RMSE of toluene-benzene and RMSE of toluene-styrene, RMSE of phenanthrene-fluoranthene-pyrene-chrysene < RMSE of phenanthrene-fluoranthene and RMSE of phenanthrene-pyrene and RMSE of phenanthrene-chrysene. This phenomenon is consistent with the results of previous studies (Zhang and Yang 2017).

Furthermore, the RMSEs shown in Fig. 6 basically increase as the number of neighborhood sample points increases, and then, they become stable. This is dependent on the soil pollution characteristics of the site and the RMSE principle. According to the computational formula for the RMSE, the value of the RMSE is greatly affected by the maximum difference between the measured value and the predicted value. For a contaminated site, the maximum error is mainly located in the extremely high value position (underestimated).

A small number of extremely high values in the site deviate from the overall trend of pollutant content. When the number of neighboring samples increases, the prediction result of the high values is greatly affected by the low values in the neighborhood. Therefore, the underestimation of the extremely high values is more significant for higher RMSEs. In addition, because the variability in the low value area, except for the high values, is relatively small, when the number of neighboring points increases to a certain number, the impact on the prediction results of the high values is small.

This result is contrary to the results of previous studies (Zhang and Yang 2017). This is due to the differences in the spatial variation characteristics of the research objects. Soil organic matter is a type of non-point source pollution, with a strong spatial continuity. Increasing the number of neighboring sample points results in the better characterization of the pollutant content of the point to be determined using

the distribution characteristics of the neighboring sample points. However, the spatial heterogeneity of the pollutants at the site is strong, and the local characteristics are clear (Wu et al. 2011; Liu et al. 2013). Increasing the number of neighboring sample points can't reveal the local characteristics of the extreme value position.

4.3 The relationships between the number of sample points and the accuracy of prediction combination with auxiliary pollutants

Reducing the number of investigation sample points reduces the accuracy of the prediction, while adding auxiliary pollutants improves the accuracy of the prediction. Therefore, theoretically, after adding auxiliary pollutants, the number of investigation sample points can be reduced to a certain extent, and the cos of the investigation can be decreased. The relationship between the number of samples removed and the accuracy of the prediction is analyzed in this section.

Because of the strong correlation between a target pollutant and its auxiliary pollutant(s), the sample points were removed based on the spatial distribution characteristics of the target pollutant. The principle of sample removal is as follows. (1) Maintain the local characteristics of the site, i.e., keep the maximum and minimum values in the data. (2) Keep the low value near the high value point to maintain the local characteristics of the site. (3) Using the group analysis method in the ArcGIS software, divide the pollutant data into 20 groups, and evenly remove points from the level with the highest number of points.

The RMSE increases as the number of sample points removed increases. When 84 sample points are removed, RMSE of toluene-benzene-styrene) ≈ RMSE of toluene based on the total number of samples. This is due to the significant clustering distribution of the toluene concentration. After removing points from the categories containing more sample points, the detail representation weakens and the RMSE gradually increases.

The cross distribution of the high and low concentrations of Cu and phenanthrene are significant. RMSE of Cu-Cd-As and RMSE of phenanthrene-fluoranthene-pyrene-chrysene initially increased and then decreased as the number of sample points removed increases. This is because at the beginning of the sample removal, the number of sample points decreased, and the spatial distribution details of the Cu and phenanthrene concentrations could not be characterized, so the RMSE increased. As the number of sample points used continued to decrease (i.e., increased removal of points), the local variation characteristics of the site were lost, resulting in a weak overall variability, easy fitting, and reduced RMSE. When the number of sample points removed were 61 and 34 for Cu and phenanthrene, respectively, RMSE of Cu-Cd-As and RMSE of phenanthrene-fluoranthene-pyrene-chrysene were close to RMSE of Cu and RMSE of phenanthrene based on the total number of samples.

Conclusions

Based on the related pollutants, the reduction in RMSEs of Cu, toluene, and phenanthrene for multivariable cokriging were 68.4%, 81.6%, and 81.2%, respectively, which are larger than the reduction in their RMSEs when they are combined with a single auxiliary variable. In addition, the reduction in RMSEs of the spatial distributions of the target pollutants combined with their auxiliary pollutants increase with increasing correlation between the auxiliary pollutants and the target pollutants. RMSE increases as the number of neighboring sample points increases, and then, it becomes stable. Furthermore, for the target pollutants combined with their related pollutants, the prediction accuracies of the high value regions are improved. In addition, with the same prediction accuracy, 34, 61 and 84 sample points can be saved by adding auxiliary pollutants for phenanthrene, Cu and toluene, respectively.The research results of this study provide important information for improving the prediction accuracy of the spatial distributions of soil pollutants and reducing remediation cost.

Declarations

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of data and materials: The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests: The authors declare that they have no competing interests.

Funding: This work was supported by the Sprout project of Beijing Academy of Science and Technology.

Authors' contributions: QPW was a major contributor in writing the manuscript. LDL is responsible for the preliminary data collation. YSC controlled the content of the whole article. ZQY and WHQ collected and analyzed samples. All authors read and approved the final manuscript.

Acknowledgements: We thank LetPub (www.letpub.com) for its linguistic assistance during the preparation of this manuscript.

References

1. Armiento G, Cremisini C, Nardi E, Pacifico R (2011) High geochemical background of potentially harmful elements in soils and sediments: implications for the remediation of contaminated sites. Chem Ecol 27:131–141. DOI:10.1080/02757540.2010.534085
2. Carlon C, Critto A, Marcomini A, Nathanail P (2001) Risk based characterisation of contaminated industrial site using multivariate and geostatistical tools. Environ Pollut 111:417–427. DOI:10.1016/s0269-7491(00)00089-0
3. Chen XB, Liu WX, Zhou YX, Qiao XY, Zhao JB, Li HF et al (2016) Analysis of HCHs and DDTs in a typical prestide contaminated site. Fresenius Environ Bull 25:5145–5150

4. Cui JL, Luo CL, Tang CWY, Chan TS, Li XD (2017) Speciation and leaching of trace metal contaminants from e-waste contaminated soils. J Hazard Mater 329:150–158. DOI:10.1016/j.jhazmat.2016.12.060
5. Ding Q, Cheng G, Wang Y, Zhuang DF (2017) Effects of natural factors on the spatial distribution of heavy metals in soils surrounding mining regions. Sci Total Environ 578:577–585. DOI:10.1016/j.scitotenv.2016.11.001
6. Dong JH, Yu M, Bian ZF, Wang Y, Di CL (2011) Geostatistical analyses of heavy metal distribution in reclaimed mine land in Xuzhou, China. Environ Earth Sci 62:127–137. DOI:10.1007/s12665-010-0507-5
7. Fang YY, Nie ZQ, Die QQ, Tian YJ, Liu F, He J et al (2017) Organochlorine pesticides in soil, air, and vegetation at and around a contaminated site in southwestern China: Concentration, transmission, and risk evaluation. Chemosphere 178:340–349. DOI:10.1016/j.chemosphere.2017.02.151
8. Fu CC, Zhang HB, Tu C, Li LZ, Luo YM (2018) Geostatistical interpolation of available copper in orchard soil as influenced by planting duration. Environ Sci Pollut Res 25:52–63. DOI:10.1007/s11356-016-7882-8
9. Gao L, Shao MG (2012) The interpolation accuracy for seven soil properties at various sampling scales on the Loess Plateau, China. J Soils Sediments 12:128–142. DOI:10.1007/s11368-011-0438-0
10. Girault F, Perrier F, Poitou C, Isambert A, Theveniaut H, Laperche V et al (2016) Effective radium concentration in topsoils contaminated by lead and zinc smelters. Sci Total Environ 566:865–876. DOI:10.1016/j.scitotenv.2016.05.007
11. Goovaerts P, Trinh HT, Demond AH, Towey T, Chang SC, Gwinn D et al (2008) Geostatistical modeling of the spatial distribution of soil dioxin in the vicinity of an incinerator. 2. Verification and calibration study. Environmental Science Technology 42:3655–3661. DOI:10.1021/es7024966
12. Gou YL, Yang SC, Cheng YJ, Song Y, Qiao PW, Li PZ et al (2019) Enhanced anoxic biodegradation of polycyclic aromatic hydrocarbons (PAHs) in aged soil pretreated by hydrogen peroxide. Chem Eng J 356:524–533. DOI:10.1016/j.cej.2018.09.059
13. Gutierrez M, Wu SS, Peebles JL (2015) Geochemical mapping of Pb- and Zn-contaminated streambed sediments in southwest Missouri, USA. J Soils Sediments 15:189–197. DOI:10.1007/s11368-014-1010-5
14. Hofmann T, Darsow A, Schafmeister MT (2010) Importance of the nugget effect in variography on modeling zinc leaching from a contaminated site using simulated annealing. J Hydrol 389:78–89. DOI:10.1016/j.jhydrol.2010.05.024
15. Huo XN, Li H, Sun DF, Zhou LD, Li BG (2010) Multi-scale spatial structure of heavy metals in agricultural soils in Beijing. Environ Monit Assess 164:605–616. DOI:10.1007/s10661-009-0916-7
16. Juang KW, Lee DY (1998) A Comparison of Three Kriging Methods Using Auxiliary Variables in Heavy-Metal Contaminated Soils. J Environ Qual 27:355–363

17. Juang KW, lee DY, Chen ZS (1996) Prediction of spatial distribution of heavy metal in contaminated soils by geostatistics: I. Effect of extreme values and semivariogram models. J Chin Chem Soc 34:560–574
18. Juang KW, Lee DY, Ellsworth TR (2001) Using rank-order geostatistics for spatial interpolation of highly skewed data in a heavy-metal contaminated site. J Environ Qual 30:894–903. DOI:10.2134/jeq2001.303894x
19. Khosravi Y, Balyani S (2019) Spatial Modeling of Mean Annual Temperature in Iran: Comparing Cokriging and Geographically Weighted Regression. Environmental Modeling Assessment 24:341–354. DOI:10.1007/s10666-018-9623-5
20. Le C, Le T, Jeong HD, Lee EB (2019) Geographic Information System–Based Framework for Estimating and Visualizing Unit Prices of Highway Work Items. Journal of Construction Engineering Management 145:04019044. DOI:10.1061/(asce)co.1943-7862.0001672
21. Li J, Heap AD (2014) Spatial interpolation methods applied in the environmental sciences: A review. Environ Model Softw 53:173–189. DOI:10.1016/j.envsoft.2013.12.008
22. Li X, Jiao W, Xiao R, Chen W, Liu W (2017) Contaminated sites in China: Countermeasures of provincial governments. J Clean Prod 147:485–496. DOI:10.1016/j.jclepro.2017.01.107
23. Li XW, Xie YF, Wang JF, Christakos G, Si JL, Zhao HN et al (2013) Influence of planting patterns on fluoroquinolone residues in the soil of an intensive vegetable cultivation area in northern China. Sci Total Environ 458–460:63–69. DOI:10.1016/j.scitotenv.2013.04.002
24. Li XY, Liu LJ, Wang YG, Luo GP, Chen X, Yang XL et al (2013) Heavy metal contamination of urban soil in an old industrial city (Shenyang) in Northeast China. Geoderma 192:50–58. DOI:10.1016/j.geoderma.2012.08.011
25. Liu G, Bi R, Wang S, Li F, Guo G (2013) The use of spatial autocorrelation analysis to identify PAHs pollution hotspots at an industrially contaminated site. Environ Monit Assess 185:9549–9558. DOI:10.1007/s10661-013-3272-6
26. Liu G, Niu JJ, Guo WJ, Zhao L, Zhang C, Wang M et al (2017) Assessment of terrain factors on the pattern and extent of soil contamination surrounding a chemical industry in Chongqing, Southwest China. Catena 156:237–243. DOI:10.1016/j.catena.2017.04.005
27. Liu G, Niu JJ, Zhang C, Guo GL (2015) Accuracy and uncertainty analysis of soil Bbf spatial distribution estimation at a coking plant-contaminated site based on normalization geostatistical technologies. Environ Sci Pollut Res 22:20121–20130. DOI:10.1007/s11356-015-5122-2
28. Lu AX, Wang JH, Qin XY, Wang KY, Han P, Zhang SZ (2012) Multivariate and geostatistical analyses of the spatial distribution and origin of heavy metals in the agricultural soils in Shunyi, Beijing, China. Sci Total Environ 425:66–74. DOI:10.1016/j.scitotenv.2012.03.003
29. Ma ZW, Chen K, Li ZY, Bi J, Huang L (2016) Heavy metals in soils and road dusts in the mining areas of Western Suzhou, China: a preliminary identification of contaminated sites. J Soils Sediments 16:204–214. DOI:10.1007/s11368-015-1208-1

30. Modis K, Papantonopoulos G, Komnitsas K, Papaodysseus K (2008) Mapping optimization based on sampling size in earth related and environmental phenomena. Stoch Env Res Risk Assess 22:83–93. DOI:10.1007/s00477-006-0096-8
31. Monaco D, Riccio A, Chianese E, Adamo P, Di Rosa S, Fagnano M (2015) Chemical characterization and spatial distribution of PAHs and heavy hydrocarbons in rural sites of Campania Region, South Italy. Environ Sci Pollut Res 22:14993–15003. DOI:10.1007/s11356-015-4733-y
32. DB11/T 811 (2011) Screening Levels for Soil Environmental. Risk Assessment of Sites
33. Paulette L, Man T, Weindorf DC, Person T (2015) Rapid assessment of soil and contaminant variability via portable x-ray fluorescence spectroscopy: Copşa Mică, Romania. Geoderma 243–244:130–140. DOI:10.1016/j.geoderma.2014.12.025
34. Qiao PW, Lei M, Yang SC, Yang J, Guo GH, Zhou XY (2018) Comparing ordinary kriging and inverse distance weighting for soil as pollution in Beijing. Environ Sci Pollut Res 25:15597–15608. DOI:10.1007/s11356-018-1552-y
35. Qiao PW, Li PZ, Cheng YJ, Wei WX, Yang SC, Lei M et al (2019) Comparison of common spatial interpolation methods for analyzing pollutant spatial distributions at contaminated sites. Environ Geochem Health 41:2709–2730. DOI:10.1007/s10653-019-00328-0
36. Rashed MN (2010) Monitoring of contaminated toxic and heavy metals, from mine tailings through age accumulation, in soil and some wild plants at Southeast Egypt. J Hazard Mater 178:739–746. DOI:10.1016/j.jhazmat.2010.01.147
37. Ren LX, Lu HW, He L, Zhang YM (2016) Characterization of monochlorobenzene contamination in soils using geostatistical interpolation and 3D visualization for agrochemical industrial sites in southeast China. Archives of Environmental Protection 42:17–24. DOI:10.1515/aep-2016-0025
38. Robinson TP, Metternicht G (2006) Testing the performance of spatial interpolation techniques for mapping soil properties. Comput Electron Agric 50:97–108. DOI:10.1016/j.compag.2005.07.003
39. Roslund MI, Gronroos M, Rantalainen AL, Jumpponen A, Romantschuk M, Parajuli A et al (2018) Half-lives of PAHs and temporal microbiota changes in commonly used urban landscaping materials. PeerJ 6:e4508. DOI:10.7717/peerj.4508
40. Saito H, Goovaerts P (2000) Geostatistical interpolation of positively skewed and censored data in a dioxin-contaminated site. Environmental Science Technology 34:4228–4235. DOI:10.1021/es991450y
41. Saito H, Goovaerts P (2002) Accounting for measurement error in uncertainty modeling and decision-making using indicator kriging and p-field simulation: application to a dioxin contaminated site. Environmetrics 13:555–567. DOI:10.1002/env.545
42. Santos-Francés F, Martinez-Graña A, Alonso Rojo P, García Sánchez A (2017) Geochemical background and baseline values determination and spatial distribution of heavy metal pollution in soils of the Andes mountain range (Cajamarca-Huancavelica, Peru). International Journal of Environmental Research Public Health 14:859. DOI:10.3390/ijerph14080859

43. Schnabel U, Tietje O (2003) Explorative data analysis of heavy metal contaminated soil using multidimensional spatial regression. Environ Geol 44:893–904. DOI:10.1007/s00254-003-0844-8
44. Shi R, Xu M, Liu A, Tian Y, Zhao Z (2017) Characteristics of PAHs in farmland soil and rainfall runoff in Tianjin, China. Environ Monit Assess 189:558. DOI:10.1007/s10661-017-6290-y
45. Tziachris P, Metaxa E, Papadopoulos F, Papadopoulou M (2017) Spatial Modelling and Prediction Assessment of Soil Iron Using Kriging Interpolation with pH as Auxiliary Information. ISPRS International Journal of Geo-Information 6:283. DOI:10.3390/ijgi6090283
46. USEPA (2014) Method 6020B: Inductively Coupled Plasma - Mass Spectrometry
47. USEPA (2018) Method 8270E: Semivolatile Organic Compounds by Gas Chromatography/Mass Spectrometry. GC/MS)
48. USEPA (2017) Method 8260D (SW-846): Volatile Organic Compounds by Gas. Chromatography-Mass Spectrometry (GC/MS)
49. Vyas VM, Tong SN, Uchrin C, Georgopoulos PG, Carter GR (2004) Geostatistical estimation of horizontal hydraulic conductivity for the Kirkwood-Cohansey aquifer. J Am Water Resour Assoc 40:187–195. DOI:10.1111/j.1752-1688.2004.tb01018.x
50. Weindorf DC, Paulette L, Man T (2013) In-situ assessment of metal contamination via portable X-ray fluorescence spectroscopy: Zlatna, Romania. Environ Pollut 182:92–100. DOI:10.1016/j.envpol.2013.07.008
51. Wu C, Wu J, Luo Y, Zhang H, Teng Y, DeGloria SD (2011) Spatial interpolation of severely skewed data with several peak values by the approach integrating kriging and triangular irregular network interpolation. Environ Earth Sci 63:1093–1103
52. Wu CF, Wu JP, Luo YM, Zhang HB, Teng Y (2008) Statistical and geoestatistical characterization of heavy metal concentrations in a contaminated area taking into account soil map units. Geoderma 144:171–179. DOI:10.1016/j.geoderma.2007.11.001
53. Wu CF, Wu JP, Luo YM, Zhang HB, Teng Y, DeGloria SD (2011) Spatial interpolation of severely skewed data with several peak values by the approach integrating kriging and triangular irregular network interpolation. Environ Earth Sci 63:1093–1103. DOI:10.1007/s12665-010-0784-z
54. Wu GZ, Kechavarzi C, Li XG, Wu SM, Pollard SJT, Sui H et al (2013) Machine learning models for predicting PAHs bioavailability in compost amended soils. Chem Eng J 223:747–754. DOI:10.1016/j.cej.2013.02.122
55. Wu J, Norvell WA, Welch RM (2006) Kriging on highly skewed data for DTPA-extractable soil Zn with auxiliary information for pH and organic carbon. Geoderma 134:187–199. DOI:10.1016/j.geoderma.2005.11.002
56. Xie YF, Chen TB, Lei M, Yang J, Guo QJ, Song B et al (2011) Spatial distribution of soil heavy metal pollution estimated by different interpolation methods: accuracy and uncertainty analysis. Chemosphere 82:468–476. DOI:10.1016/j.chemosphere.2010.09.053
57. Yang QY, Luo WQ, Jiang ZC, Li WJ, Yuan DX (2016) Improve the prediction of soil bulk density by cokriging with predicted soil water content as auxiliary variable. J Soils Sediments 16:77–84.

DOI:10.1007/s11368-015-1193-4

58. Yang SC, Gou YL, Song Y, Li PZ (2018) Enhanced anoxic biodegradation of polycyclic aromatic hydrocarbons (PAHs) in a highly contaminated aged soil using nitrate and soil microbes. Environ Earth Sci 77. DOI:10.1007/s12665-018-7629-6
59. Zhang B, Yang Y (2017) Spatiotemporal modeling and prediction of soil heavy metals based on spatiotemporal cokriging. Sci Rep 7:10. DOI:10.1038/s41598-017-17018-5
60. Zhen JC, Pei T, Xie SY (2019) Kriging methods with auxiliary nighttime lights data to detect potentially toxic metals concentrations in soil. Sci Total Environ 659:363–371. DOI:10.1016/j.scitotenv.2018.12.330
61. Zhou J, Feng K, Li Y, Zhou Y (2016) Factorial Kriging analysis and sources of heavy metals in soils of different land-use types in the Yangtze River Delta of Eastern China. Environ Sci Pollut Res 23:14957–14967. DOI:10.1007/s11356-016-6619-z

Figures

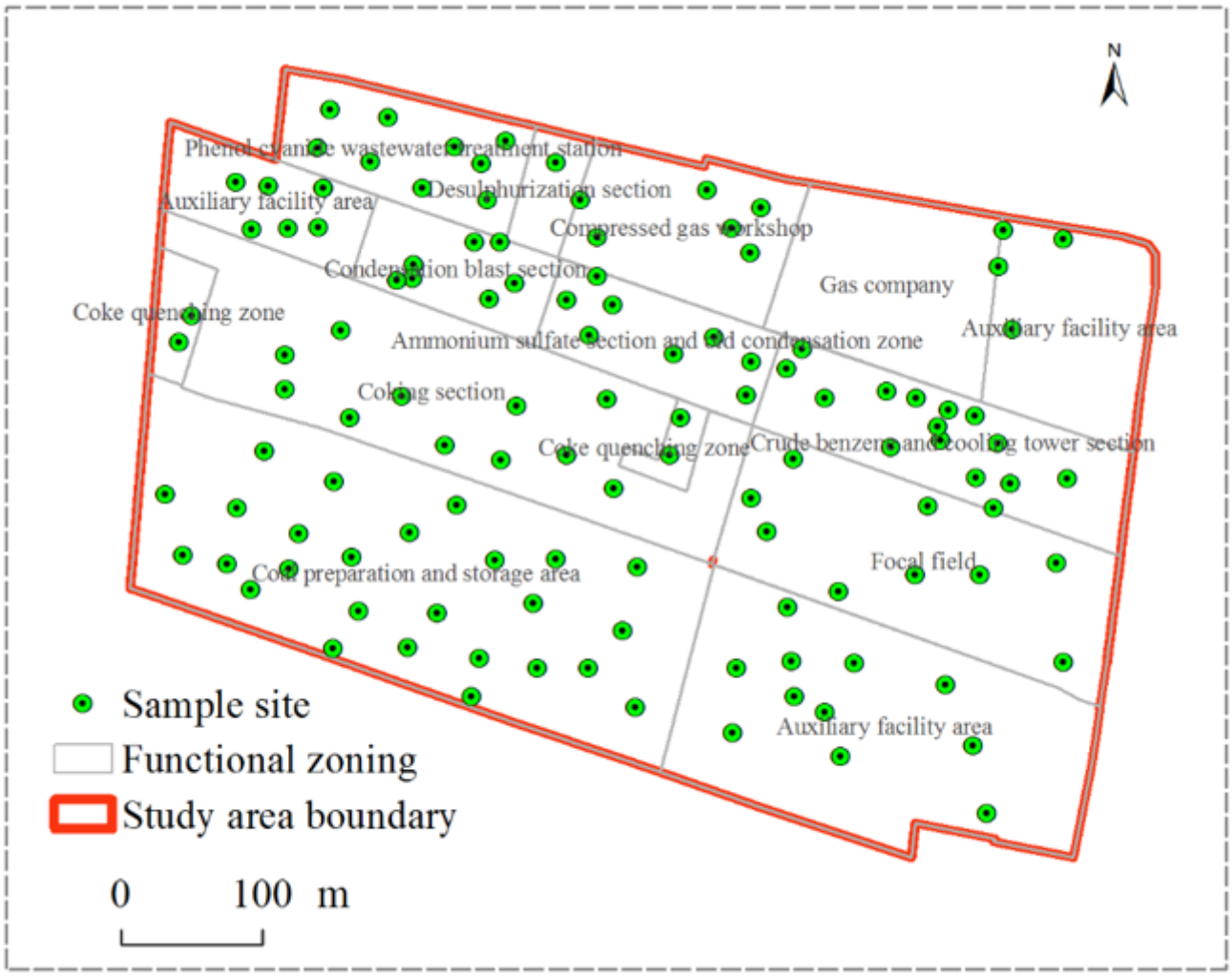

Figure 1

The sample locations

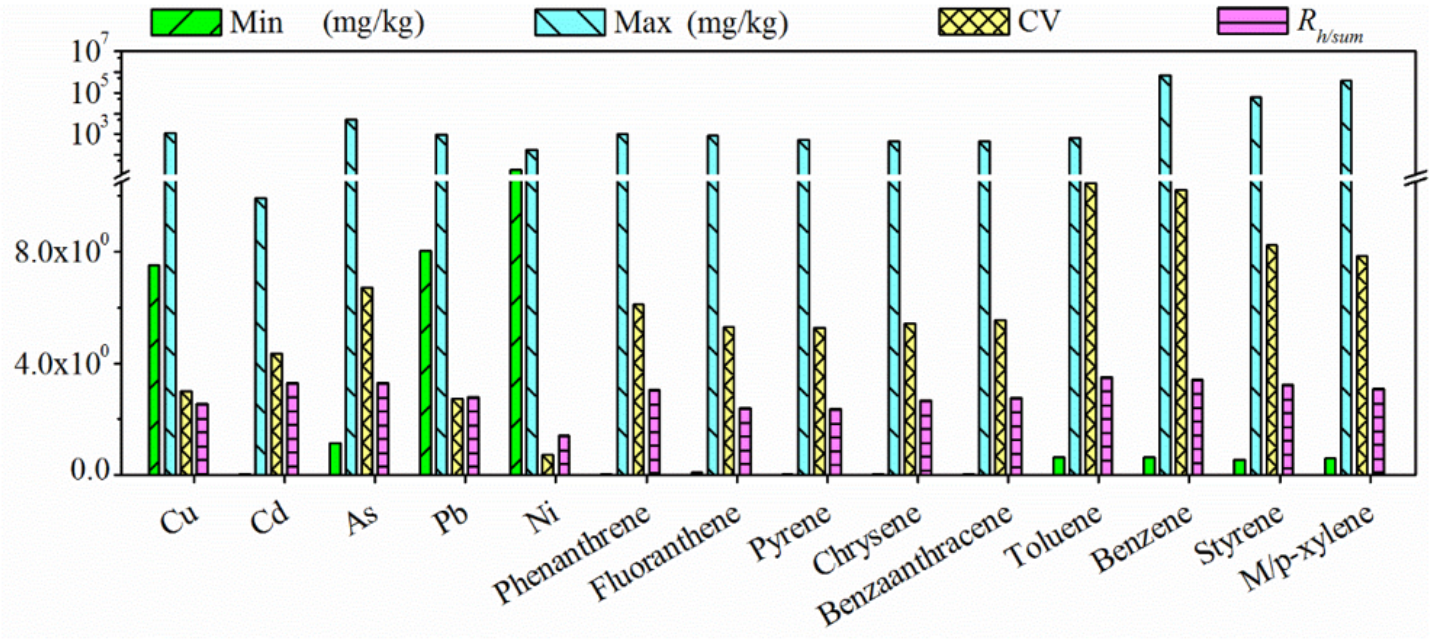

Figure 2

Descriptive statistics for Cu, phenanthrene, and toluene and their related pollutants

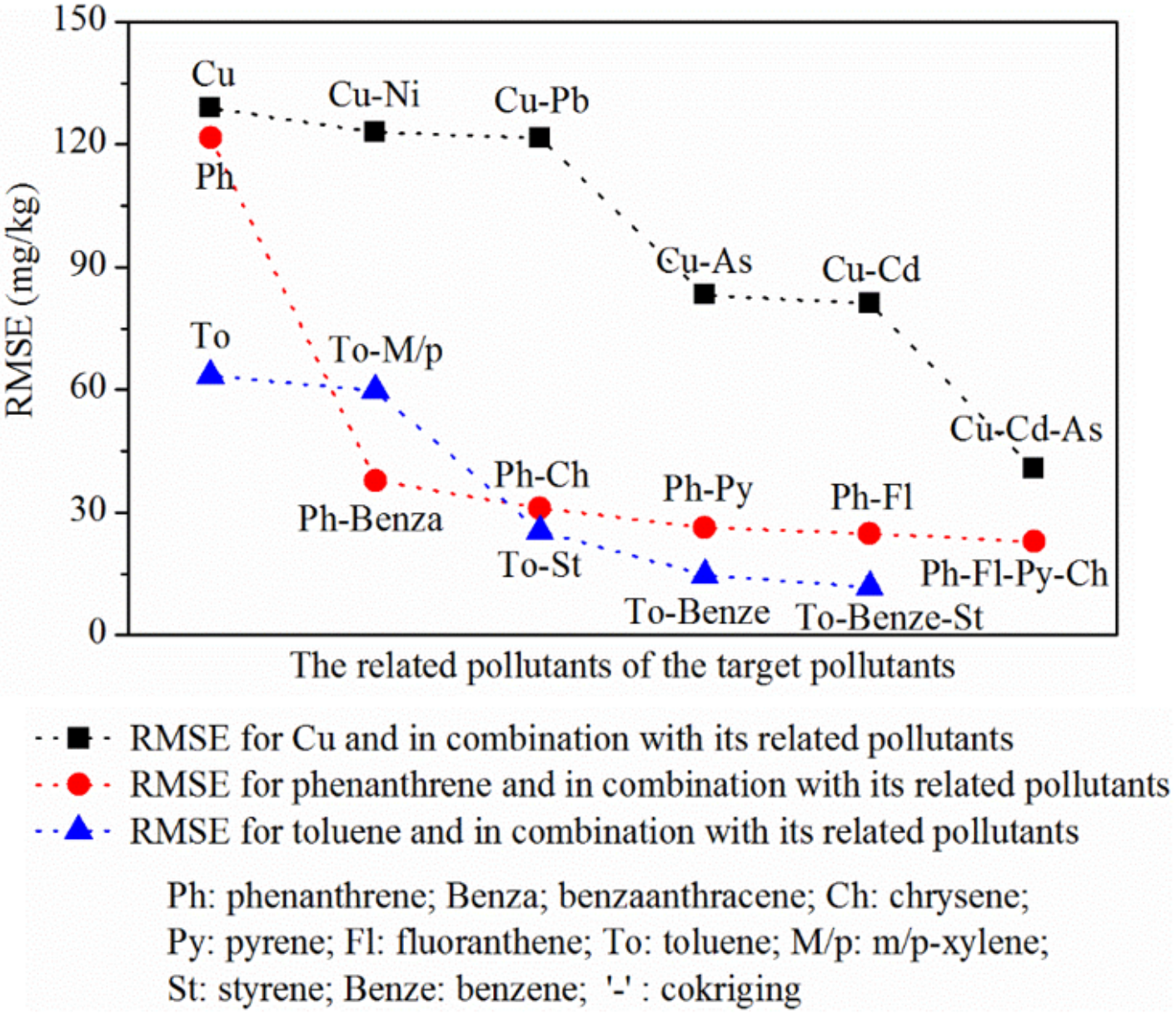

Figure 3

Cross-validation results of the interpolations of Cu, phenanthrene, and toluene and in combination with their related pollutants

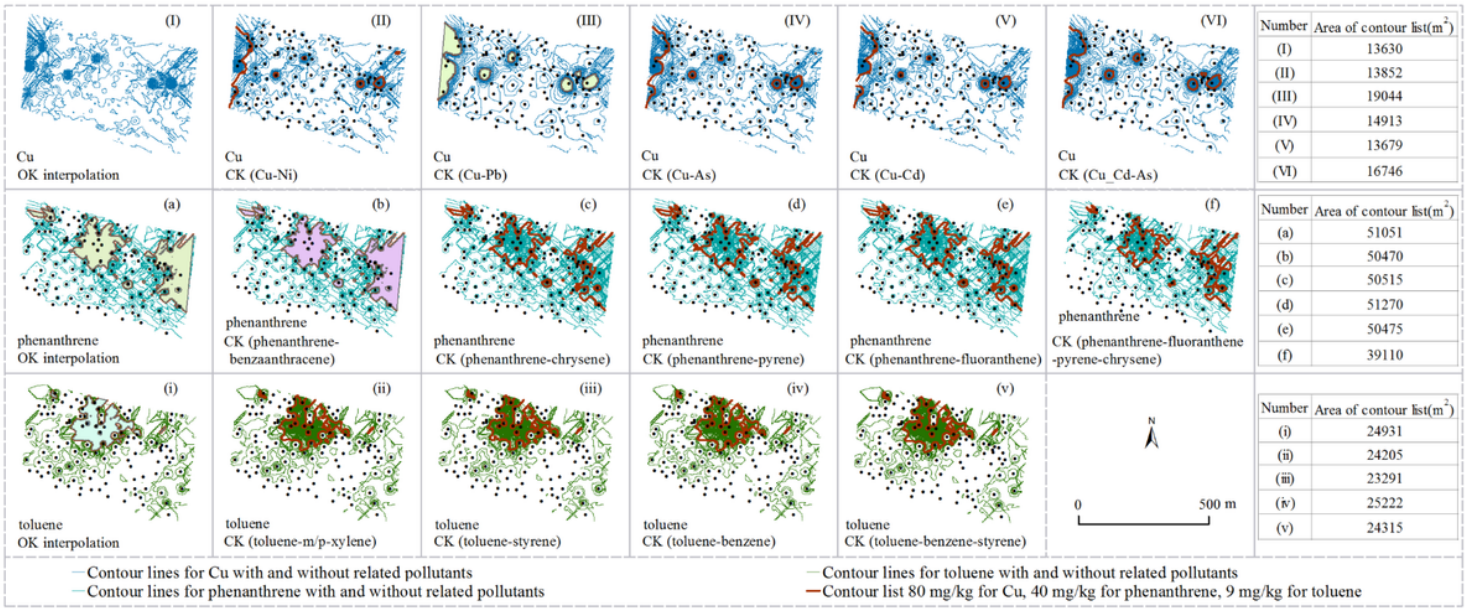

Figure 4

The contour lines (1) for Cu predicted using the OK interpolation method and (2–6) for Cu combined with its related pollutants using the CK interpolation method. The contour lines for (a) phenanthrene predicted using the OK interpolation method and (b–f) for phenanthrene combined with its related pollutants using the CK interpolation method. The contour lines (i) for toluene predicted using the OK interpolation method and (ii–v) for toluene combined with its related pollutants using the CK interpolation method.

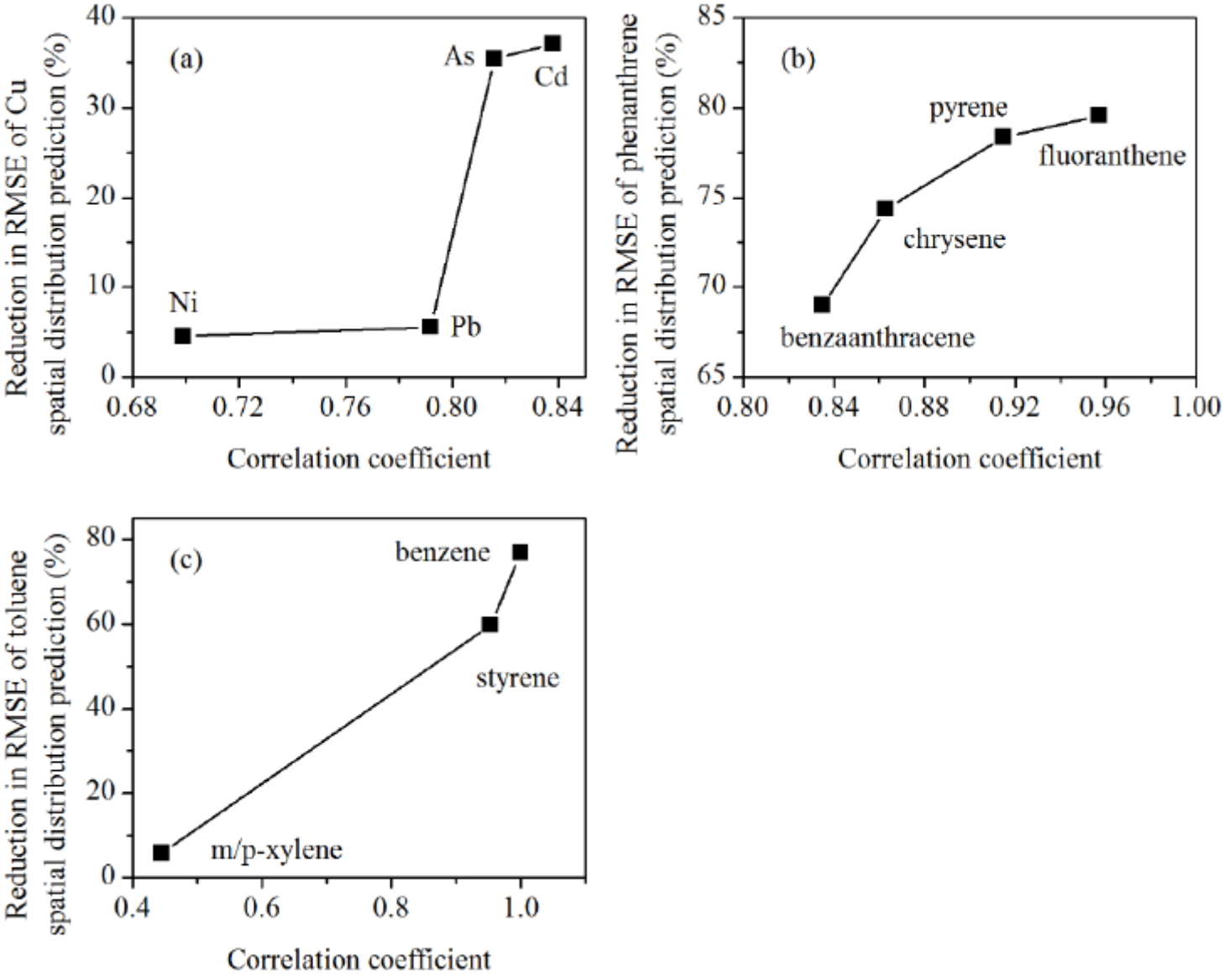

Figure 5

Curve showing the change in the correlation coefficient as the root mean square error (RMSE) of the spatial distribution prediction result of the target pollutants decreases. (a) Cu, (b) Phenanthrene, (c) Toluene

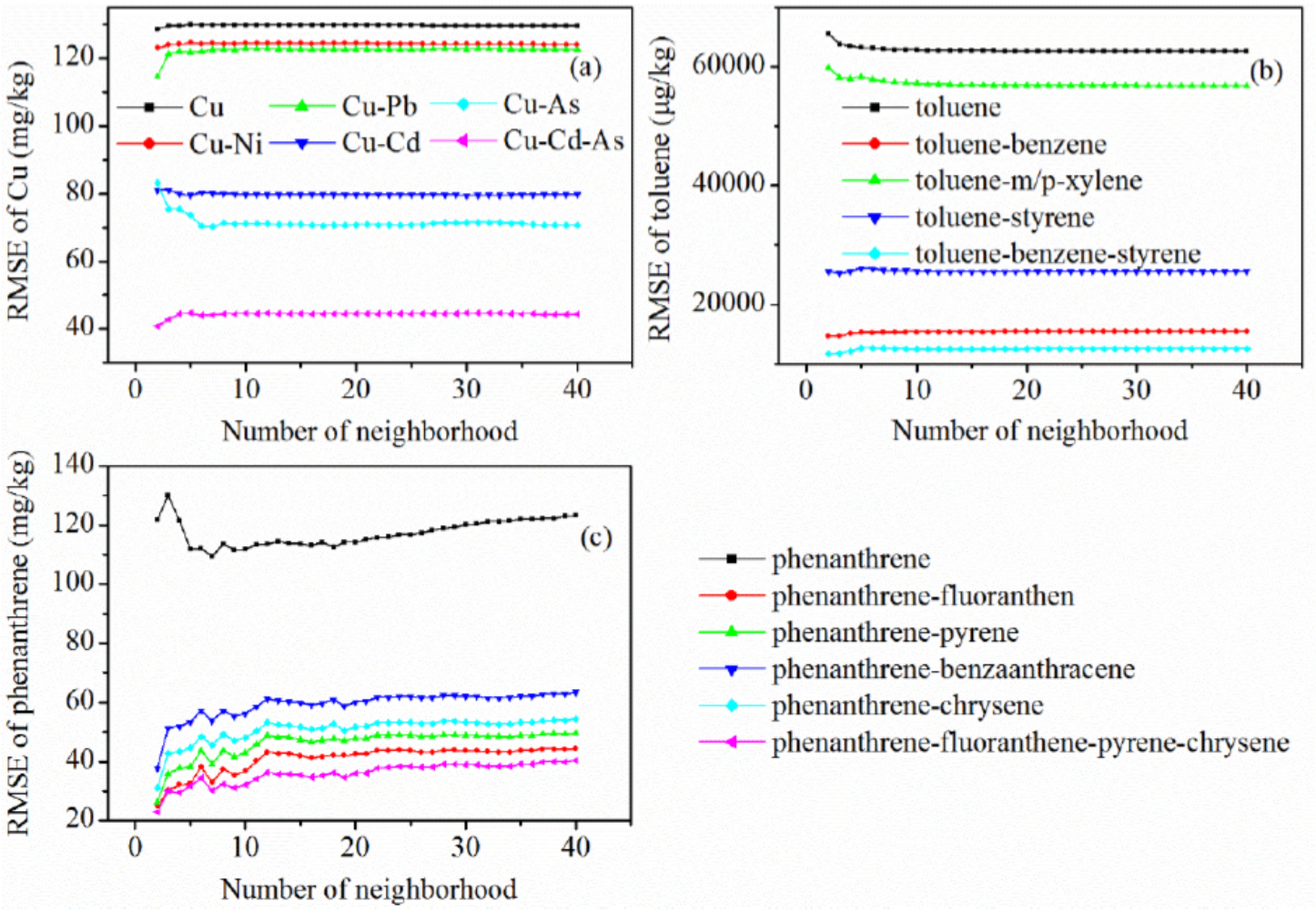

Figure 6

RMSEs of (a) Cu, (b) toluene, and (c) phenanthrene with different numbers of neighboring points and different combinations of auxiliary pollutants

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Graphicalabstract20200831.tif