Stochastic Environmental Research and Risk Assessment
https://doi.org/10.1007/s00477-018-1641-y

ORIGINAL PAPER

Improving the spatial prediction of soil Zn by converting outliers into soft data for BME method

Chu-tian Zhang[1] · Yong Yang[2,3]

Abstract

Understanding the spatial patterns of heavy metals is important for the protection and remediation of urban soil. Considering that the conventional Geostatistical methods, such as ordinary kriging (OK), are sensitive to dataset outliers, this study converted the identified outliers into a discrete probability density function (PDF). Then, the PDF was used as soft data in the Bayesian maximum entropy (BME) framework to perform a spatial prediction of soil Zn contents in Wuhan City, Central China. By using OK as the reference method, the BME framework was found to produce an overall further accurate prediction, and the PDF of BME predictions was further informative and close to the observed Zn concentrations. An improved BME performance can be expected if soft data with high quality are provided. The BME is a promising method in environmental science, where the so-called outliers that probably carry important information are common.

Keywords Bayesian maximum entropy · Soil Zn contents · Outliers · Discrete probability density function · Soft data

1 Introduction

Soil contamination by heavy metals has been receiving increasing attention worldwide in recent years, especially in urban soils due to rapid industrialization and urbanization. Heavy metals in soil can be hardly eliminated due to their nonbiodegradable nature and long biological half-life (Guo et al. 2012). Elevated concentrations of heavy metals in urban soils may threaten the health of citizens, especially the children, because these metals can be easily transferred into human bodies through polluted food ingestion or direct contact (Benhaddya and Hadjel 2014). Moreover, the polluted urban soils may enter the atmosphere as dust (Meza-Figueroa et al. 2007) and affect water quality due to surface runoff and soil erosion (Helmreich et al. 2010), thereby being further carried into sensitive environments. Therefore, accurate knowledge of the spatial pattern of heavy metals is vital for the risk assessment, protection, and remediation of urban soil.

Geostatistical methods, mainly including the various kriging techniques, have been widely applied in soil sciences and are effective for quantifying the spatial features of soil attributes, such as soil heavy metals (Webster and Oliver 2007). A properly selected variogram model, which depicts the spatial structure of the variable under study, is the key for the successful application of kriging techniques. However, the existence of outliers will greatly affect the variogram form and cause the erratic behavior of the variogram model (McGrath and Zhang 2003). A popular way of managing outliers is to compute and fit the robust variogram models, which are less sensitive to outliers (Cressie and Hawkins 1980; Lark 2000). In fact, data for the concentrations of heavy metals often contain "outliers" probably due to potential contamination (Zhang et al. 2009). The "outliers" are supposed to come from a second process (e.g., contamination) in the robust variogram

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00477-018-1641-y) contains supplementary material, which is available to authorized users.

✉ Yong Yang
yangyong@mail.hzau.edu.cn

[1] College of Natural Resources and Environment, Northwest A&F University, Yangling 712100, China

[2] College of Resources and Environment, Huazhong Agricultural University, Wuhan 430070, China

[3] Key Laboratory of Arable Land Conservation (Middle and Lower Reaches of Yangtze River), Ministry of Agriculture, Wuhan, China

Published online: 02 January 2019

models. This method seeks to weaken the influences of the outliers, but these outliers probably carry critical information that should be processed carefully and directly.

Bayesian maximum entropy (BME), which was introduced in the 1990s (Christakos 1990, 2000), provides a physically meaningful and formally rigorous framework for synthesizing multisource data and knowledge, thereby improving modeling and spatial prediction (Gao et al. 2014; Savelieva et al. 2005). A major characteristic of the BME is that it utilizes data with uncertainty (hereinafter, the soft data) in a flexible manner. For instance, setting the prediction intervals as interval soft data is a common practice, whereas the prediction intervals are often derived from the empirical relationships between the target variable and highly correlated auxiliary information (Douaik et al. 2005; Gao et al. 2014). Furthermore, the probability distribution of various forms can be constructed from a time series with missing values and duplicated observations and then be used as probability soft data (Reyes and Serre 2014; Savelieva et al. 2005). Puangthongthub et al. (2007) regarded high outliers above the 99 percentile value of measured PM_{10} concentrations as soft data. These outliers were described by a Gaussian probability density function (PDF) with a standard deviation (SD) that is equivalent to the sampling method accuracy or sufficiently large to cover the 99 percentile value.

In this study, soil Zn was selected as the experimental trace element due to its ubiquity in urban soils and toxicity to human health. Dataset outliers of soil Zn concentrations were initially identified. Then, the discrete probability soft data were constructed on the basis of the detected outliers and used for BME modeling. The main objective of this study was to test the capability of the BME method in the spatial prediction of soil heavy metals with the existence of outliers. For comparison, ordinary kriging (OK) based on the robust variogram model was also applied to the same dataset.

2 Materials and methods

2.1 Study area

Wuhan City, which is the capital of Hubei Province and the largest city in central China, is located in the middle reach of Yangtze River. This city had a population of approximately 10.22 million by the end of 2013 (Wuhan Bureau of Statistics 2013). Wuhan is one of China's four major comprehensive transportation hubs with heavy traffic. Wuhan is also an important industrial city that hosts numerous ferrous smelters, metal work plants, and equipment-manufacturing plants. Hundreds of lakes with various sizes and several large rivers exist within the administrative borders of the city, which account for nearly a quarter of the total city area (Wuhan Government webpage, http://www.wh.gov.cn). In this study, an area of 1016 km^2, which covers the seven core urban districts and part of Caidian and Jiangxia districts, was selected as the study area (Fig. 1).

2.2 Sampling and analysis of soil Zn concentrations

The study area was divided into grid cells of 1 km × 1 km for sampling. A total of 467 topsoil samples (0–20 cm depth) were collected inside these grid cells in November 2013. To acquire further qualified data in a short time with less expense, all 467 soil samples were analyzed by X-ray fluorescence (XRF, Niton XL2 600, Thermo Scientific, USA). Then, 150 of the soil samples were carefully selected and analyzed through inductively coupled plasma atomic emission spectrometry (ICP-AES, VISTA-MPX, VARIAN, USA). For XRF measurements, each soil sample was blank corrected and detected for at least 90 s to ensure data accuracy. For ICP measurements, three duplications were performed for each soil sample, and blank samples were also analyzed. To integrate soil Zn data obtained by ICP-AES and XRF, a linear regression model (the R-square value and correlation coefficient of the model were 0.895 and 0.946, respectively) was established on the basis of pairwise concentrations of soil Zn at the 150 selected sample sites. This model was used to rescale the XRF data into those obtained by ICP-AES. ICP-AES and XRF failed to measure the soil Zn concentrations at two soil samples; therefore, the concentrations were discarded. Finally, a total of 465 measurements of soil Zn were adopted in this study. Figure 1 plots their spatial distribution. All soil samples were analyzed at the Key Laboratory of Arable

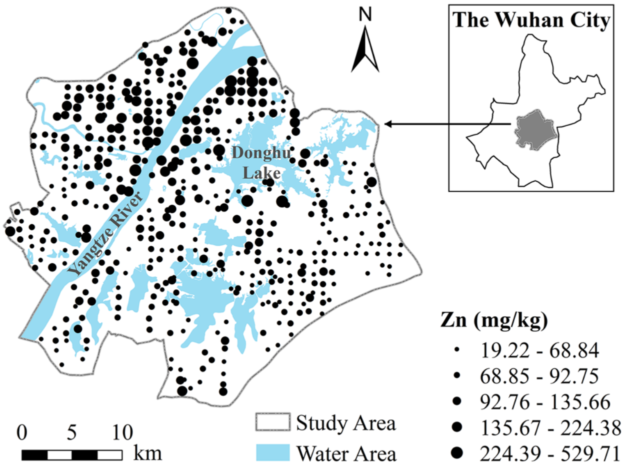


Fig. 1 Study area and spatial locations of measured Zn concentrations in soils of Wuhan city

Land Conservation (Middle and Lower Reaches of Yangtze River), Ministry of Agriculture of China. Thus, all Zn data (465 measurements) used in the following analysis can be considered accurate and free from error values. Additional details about the sampling and chemical analysis of soil heavy metals in this study area can be found in Zhang et al. (2015).

2.3 Global outlier identification

Histogram is a simple and frequently used method for demonstrating dataset distribution. This method can also provide information about outliers. As the bins (class intervals) used to count the frequencies are equally distributed, the outliers are located far away from majority of the values and thus can be visually identified (Zhang et al. 2009).

A boxplot is a nonparametric method for depicting dataset distribution through quartiles, which include the lower (25th percentile), median, and upper (75th percentile) quartiles. The bottom and top of the box are called the hinges, and the length of the box is the interquartile range (IQR). Outliers identified by boxplot consist of "mild outliers" and "extreme outliers"; the former are those values located between 1.5 × IQR from the hinges, and the latter are located beyond 3 × IQR from the hinges (Zhang et al. 2009). Only the "extreme outliers" from the boxplot were adopted in this study to retain as much information regarding the spatial variability of soil Zn as possible from the original dataset.

2.4 Spatial outlier identification and prediction by OK method

OK is adopted to identify the spatial outliers, and it is one of the most frequently used geostatistical methods, which aims to provide the best linear unbiased predictions of regional variables on the basis of intrinsic assumption (Matheron 1963; Olea 2006; Oliver and Webster 2014). The prediction of OK at an unvisited location $\boldsymbol{s}_0$ is calculated by Eq. (1).

$$\hat{z}_{OK}(\boldsymbol{s}_0) = \boldsymbol{\lambda}(\boldsymbol{s}_0)^T \boldsymbol{z} \tag{1}$$

where $\boldsymbol{\lambda}(\boldsymbol{s}_k)$ is an $N \times 1$ vector holding the weights assigned to the N observed data values $\boldsymbol{z}$ (soil Zn concentrations in this study). The prediction error variance can be easily derived by Eq. (2).

$$\sigma^2_{OK}(\boldsymbol{s}_0) = c_0 + c_1 - \boldsymbol{C}(\boldsymbol{s}_0)^T \boldsymbol{\lambda}(\boldsymbol{s}_0) \tag{2}$$

Here, $\boldsymbol{C}(\boldsymbol{s}_0)$ represents the $N \times 1$ vector of covariance values between $\boldsymbol{s}_0$ and N data points. A single covariance $C(h)$ between a certain pair of points with their Euclidean distance being h is typically calculated from the variogram models $\gamma(h)$ with their relationship as $C(h) = c - \gamma(h)$. c denotes the so-called sill parameter and can be divided into nugget (c_0) and partial sill (c_1), such that $c = c_0 + c_1$. Three commonly used variogram models, i.e., exponential, Gaussian, and spherical models, are defined by Eqs. (3), (4), and (5), respectively.

$$\gamma(h) = c_0 + c_1(1 - e^{-3h/r}) \tag{3}$$

$$\gamma(h) = c_0 + c_1(1 - e^{-3h^2/r^2}) \tag{4}$$

$$\gamma(h) = \begin{cases} c_0 + c_1\left(\dfrac{3h}{2r} - \dfrac{h^3}{2r^3}\right) & (h < r) \\ c_0 + c_1 & (h \geq r) \end{cases} \tag{5}$$

where r refers to the range parameter.

As previously mentioned, the existence of the outliers will greatly affect the variogram model; thus, the estimated and observed values at the spatial locations of outliers tend to differ considerably, and the prediction error variance is also likely to be underestimated (Meklit et al. 2009). In this situation, the standardized prediction error $\varepsilon_s(s_0)$ (abbreviated as StdP_Err) can be used as the criterion to determine outliers, which is defined by Eq. (6).

$$\varepsilon_s(s_0) = \frac{\hat{z}_{OK}(s_0) - z_{OK}(s_0)}{\sigma_{OK}(s_0)} \tag{6}$$

The leave-one-out cross-validation of OK was used to obtain $\hat{z}_{OK}(s_0)$ and $\sigma_{OK}(s_0)$ in Eq. (6). For the spatial outliers, their corresponding absolute values of $\varepsilon_s(s_0)$ will be large. In this study, the high-value spatial outliers were identified if $\varepsilon_s(s_0)$ is smaller than $-$ 1.96 (Meklit et al. 2009; Zhang et al. 2009).

In practice, the sample variogram, to which the theoretical variogram model is fitted, is calculated. In this study, two types of sample variogram were used for different purposes. One is the most widely used estimator proposed by Matheron (1963), which was denoted as $\hat{\gamma}_M(h)$ and used for identifying spatial outliers. The other is the robust variogram proposed by Cressie and Hawkins (1980), which was denoted as $\hat{\gamma}_{CH}(h)$ and applied to the validation and prediction of OK method with the existence of identified outliers. $\hat{\gamma}_M(h)$ and $\hat{\gamma}_{CH}(h)$ (Lark 2000) are defined by Eqs. (7) and (8), respectively.

$$\hat{\gamma}_M(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [z(s_i) - z(s_i + h)]^2 \tag{7}$$

$$\hat{\gamma}_{CH}(h) = \frac{\left\{\frac{1}{N(h)} \sum_{i=1}^{N(h)} |z(s_i) - z(s_i + h)|^{0.5}\right\}^4}{0.914 + \frac{0.988}{N(h)} + \frac{0.09}{N^2(h)}} \tag{8}$$

where $N(h)$ pairs of observations among the available data are separated by lag h.

Several other methods, such as local indicators of spatial association (LISA), geographically weighted regression (GWR), and probabilistic spatiotemporal approach based on a spatial regression test (SRT-PS), can be applied to identify the spatial outliers. However, LISA is sensitive to its parameters, such as bandwidth, GWR requires an extra element as the independent variable for outlier identification (Zhang et al. 2009), and SRT-PS is suitable for spatiotemporal dataset (Xu et al. 2014). A comparison of the performances of these methods in identifying outliers is beyond the scope of this study; thus, only the OK method was adopted.

2.5 BME method

The BME framework provides a systematic and rigorous approach for incorporating various physical knowledge bases (hereinafter denoted as KB). In particular, BME considers two types of KB: (a) the general KB, K_G, which includes physical laws and statistical moments; and (b) the site-specific KB, K_S, which consists of exact numerical values across space (hard data) and soft data (Christakos 1990, 2000; Christakos and Li 1998).

The BME approach includes three main stages of synthesizing and processing K_G and K_S, as follows:

(a) *Prior stage* Considering the vector of spatial random variables $\boldsymbol{Z}_{map} = (\boldsymbol{Z}_{hard}, \boldsymbol{Z}_{soft}, Z_0)$, $\boldsymbol{z}_{hard}$ and $\boldsymbol{z}_{soft}$ denote the data values at the hard and soft data points, respectively, and z_0 is the unknown value at a prediction location, such that $\boldsymbol{z}_{map} = (\boldsymbol{z}_{hard}, \boldsymbol{z}_{soft}, z_0)$ refers to the values at all mapping points $\boldsymbol{s}_{map} = (\boldsymbol{s}_{hard}, \boldsymbol{s}_{soft}, s_0)$, where $\boldsymbol{s}_{hard}$ and $\boldsymbol{s}_{soft}$ correspond to the hard and soft data points, respectively. This stage aims to derive the multivariate prior PDF, $f_G(\boldsymbol{z}_{map})$, which accounts for the maximum amount of information provided by K_G. The BME method borrows the concept of Shannon entropy to relate K_G with $f_G(\boldsymbol{z}_{map})$ as

$$\overline{Info_G(\boldsymbol{Z}_{map})} = -\int f_G(\boldsymbol{z}_{map}) \ln f_G(\boldsymbol{z}_{map}) d\boldsymbol{z}_{map} \tag{9}$$

where the operator $\overline{(\cdot)}$ is used to obtain the expectation value. The general knowledge is explicitly expressed as $g_\alpha(\boldsymbol{z}_{map})$, a set of functions of $\boldsymbol{z}_{map}$, such as the mean and covariance function. The specific form of $f_G(\boldsymbol{z}_{map})$ can be derived by maximizing the object function of the Lagrange multiplier method by introducing the Lagrange multiplier μ_α, as expressed in Eq. (10).

$$\begin{aligned} L[f_G(\boldsymbol{z}_{map})] = &-\int f_G(\boldsymbol{z}_{map}) \ln f_G(\boldsymbol{z}_{map}) d\boldsymbol{z}_{map} \\ &- \sum_{\alpha=0}^{N_c} \mu_\alpha \left(\int g_\alpha f_G(\boldsymbol{z}_{map}) d\boldsymbol{z}_{map} - \overline{g_\alpha} \right) \end{aligned} \tag{10}$$

(b) *Meta-prior stage* K_S is organized in a proper way, especially for $\boldsymbol{z}_{soft}$. The two typical forms of $\boldsymbol{z}_{soft}$ are the interval and probability data (Christakos and Li 1998). In this study, discrete probability data were used as soft data, i.e., certain probability values ϑ_i are known and constant in each interval, as defined by Eq. (11).

$$\begin{aligned} z_{soft} = \{ &z_i; P(z_i \in I_j) = \vartheta_j,\ i = N+1, \\ &\ldots, N_{map} - 1,\ j = 1, \ldots, N_{bins} \} \end{aligned} \tag{11}$$

The 24 identified outliers of soil Zn concentrations were submitted to the *hist* function (a generic function from the basic R package *graphics*); then, a histogram was automatically established with $N_{bins} = 9$ (Eq. (11) and Fig. 5). Thus, this single histogram is a representation of the probability density of soil Zn concentrations for all 24 outliers (the area of each bin equals ϑ_i) and can be regarded as the PDF of soft data, which is denoted as $f_S(\boldsymbol{z}_{soft})$.

(c) *Integration (posterior) stage* The prior PDF $f_G(\boldsymbol{z}_{map})$ is updated into the posterior PDF $f_K(z_0)$ that characterizes Z_0 through the operational Bayesian conditioning rule (Christakos 2000), considering the physical KB. In particular, $f_K(z_0)$ based on probability data can be expressed by Eq. (12).

$$f_K(z_0) = \frac{\int f_G(z_0, \boldsymbol{z}_{data}) f_S(\boldsymbol{z}_{soft}) d\boldsymbol{z}_{soft}}{\int f_G(\boldsymbol{z}_{data}) f_S(\boldsymbol{z}_{soft}) d\boldsymbol{z}_{soft}} \tag{12}$$

$f_K(z_0)$ provides a sufficient stochastic description of the soil Zn concentration at $\boldsymbol{s}_0$. Various nonlinear estimators can be easily derived from Eq. (12). The mean estimator was adopted in this study and defined as Eq. (13) with the prediction error variance defined by Eq. (14).

$$\hat{z}_{BME}(\boldsymbol{s}_0) = \int z_0 f_K(z_0) dz_0 \tag{13}$$

$$\sigma^2_{BME}(\boldsymbol{s}_0) = \int (z_0 - \hat{z}_0)^2 f_K(z_0) dz_0 \tag{14}$$

In this study, the soil Zn distribution was represented by the spatial random field $Zn(\boldsymbol{s}_{map}) = \overline{Zn(\boldsymbol{s}_{map})} + \varepsilon(\boldsymbol{s}_{map})$, where $\overline{Zn(\boldsymbol{s}_{map})}$ is the spatial trend of Zn concentrations in the entire study area, and $\varepsilon(\boldsymbol{s}_{map})$ indicates the spatially homogeneous residuals with zero mean (the local spatial mean is also implicitly modeled when calculating kriging weights in Eq. (1)). The software adopted in this study, namely, SEKS–GUI (Yu et al. 2007), calculated $\overline{Zn(\boldsymbol{s}_{map})}$ not only on the basis of hard data but also the soft data derived from outliers. This method is a reasonable compromise for retaining the information in outliers. In particular, the histogram expectation (also $f_S(z_{soft})$) was used to produce hard value approximations at $\boldsymbol{s}_{soft}$, and then smoothing moving window with a Gaussian kernel was applied to the hard data. The approximated hard value from soft data. $\varepsilon(\boldsymbol{s}_{hard})$, to which the theoretical covariance model was fitted, was thus obtained by subtracting the trend from $\boldsymbol{z}_{data}$ at $\boldsymbol{s}_{data}$. Therefore, the form of $f_G(z_{map})$ in this study was solved by Eq. (10) on the basis of K_G, which mainly consisted of the mean trend of Zn concentration and covariance function of Zn residuals. Finally, the trend value at s_0 was added back to the estimated value of $\varepsilon(s_0)$ on the basis of Eq. (14) to acquire the final estimation of the BME method.

2.6 Validation and comparison criteria

The outliers were identified through graphic methods (histogram and boxplot) and OK cross-validation, and then the Zn dataset without outliers was spatially randomly divided into two parts: the hard dataset with 300 data points and the validation set with 141 data points. OK and BME methods were used to predict the Zn concentrations at the validation sites, which provided pairs of predicted–observed soil Zn values. Two commonly used quantitative criteria were calculated from these pairs of values, namely, the mean error (ME) and mean squared error (MSE), where the error means the difference between the predicted and observed Zn concentrations (the estimates minus the measurements). The MSE can be divided further into three components, which represent different aspects of the discrepancy between the estimates and measurements (Douaik et al. 2005), as denoted by Eq. (15).

$$MSE = SB + SDSD + LCS \tag{15}$$

where SB, $SDSD$, and LCS are defined by Eqs. (16), (17), and (18), respectively.

$$SB = \left(\bar{\hat{z}} - \bar{z}\right)^2 \tag{16}$$

$$SDSD = \left(\sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\hat{z}_i - \bar{\hat{z}}\right)} - \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\hat{z}_i - \bar{z}\right)}\right)^2 \tag{17}$$

$$LCS = 2\sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\hat{z}_i - \bar{\hat{z}}\right)} \cdot \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\hat{z}_i - \bar{z}\right)} \cdot (1 - r) \tag{18}$$

where $\bar{\hat{z}}$ and $\bar{z}$ represent the means of the predicted and observed values at n validation sites and r represent the Pearson correlation coefficient between the predicted and observed Zn values.

Overall, an improved approach with relatively further accurate predictions should place the ME close to zero, and the MSE must be as small as possible. In detail, SB is the square of bias (ME^2), SDSD denoted the difference in the degree of dispersion between the predicted and observed values, and LCS refers to the lack of positive correlation $(1 - r)$ weighted by the SDs. A large value of SDSD and LCS indicates that the model did not estimate the magnitude and degree of fluctuation among the measurements, respectively (Douaik et al. 2005).

The OK method with robust variogram model was applied in the R package *georob*. Whereas the OK cross-validation to identify the spatial outliers was implemented in ArcGIS (Version 10.0, ESRI Inc., USA). The power parameters for Box-Cox transformation was estimated in package *forecast* with the function *BoxCox.lambda*. The BME method was adopted with SEKS-GUI v1.0.8 (Yu et al. 2007). All maps were exported from ArcGIS, and other figures were plotted in OriginPro (Version 9.0, OriginLab Corporation, USA).

3 Results and discussion

3.1 Global outliers identified by graphic methods

Figure 2 shows the histogram and basic statistics for Zn concentrations in Wuhan City. The mean Zn value (88.07 mg/kg) was higher than the corresponding natural background value in Hubei, which is 83.6 mg/kg (China National Environmental Monitoring Centre 1990). Eight soil samples had Zn measurements above the corresponding risk screening value for soil contamination of agriculture land (250 mg/kg) (Ministry of Ecology and

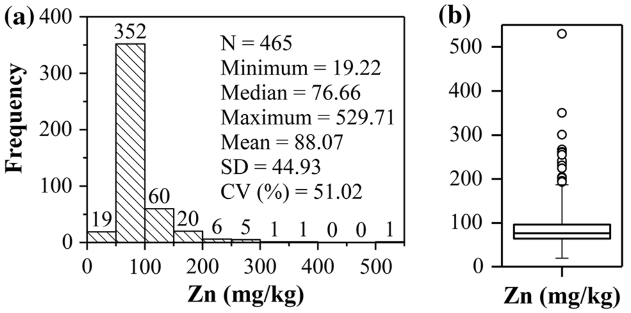


Fig. 2 Histograms and boxplots for Zn concentration in soils of Wuhan City. CV refers to the coefficient of variation. The circles represent the 16 identified global outliers

Environment of the People's Republic of China 2018). This result showed that Zn concentrations in soils of Wuhan City were elevated. The large differences between the maximum and median implied that potential high-value outliers existed. Indeed, a positively skewed distribution with a long tail extending toward the high-value side could be observed from the raw data of soil Zn concentrations (Fig. 2). Two broken bins with zero frequencies (400–450 and 450–500 mg/kg), which implied the existence of global outliers, were also observed. However, the histogram shape would be considerably different if the break values changed. On the contrary, the boxplots provide clear criteria for identifying outliers in raw data (Fig. 2). A total of 16 high-value "extreme outliers" were identified.

3.2 Spatial outliers identified through kriging method

Prior to the establishment of the variogram model, a Box–Cox transformation with a power parameter of − 0.5 was used for Zn concentrations to solve the highly positive skewed problem of raw data. A Gaussian model was used to quantify the spatial structure of the entire Zn dataset. Figure 3 shows the related parameters. The relatively high nugget-to-sill ratio (0.63 = 0.00116/0.00185) indicates that the existence of the spatial outliers will weaken the spatial continuity of Zn concentrations in Wuhan City.

The modeled variogram in Fig. 3 was used for kriging cross-validation analyses. Figure 4 presents the spatial outliers identified using the kriging method. A total of 22 high-value outliers were found, whereas only one low-value outlier was identified, indicating that the kriging method was effective in high-value outlier identification due to its smoothing effect (Zhang et al. 2009). Fourteen high-value spatial outliers were collocated with those from global outliers identified by boxplot. The collocated outliers were mainly located along the Yangtze River and Donghu Lake, thereby depicting potential soil contamination of Zn at these sites.

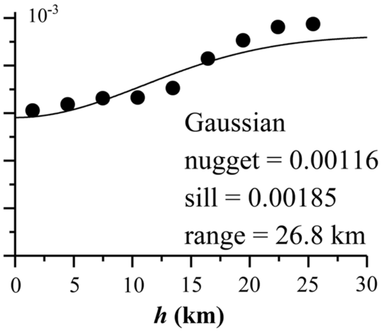


Fig. 3 Variogram of Box–Cox-transformed Zn concentration in soils of Wuhan City

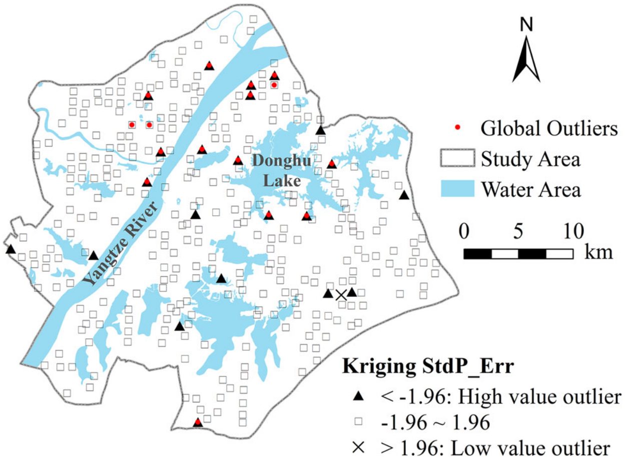


Fig. 4 Spatial outliers identified by cross-validation of kriging method. The global outliers are also plotted (red solid circles)

A total of 24 outliers (16 global outliers and 22 spatial outliers, with 14 duplicated outliers removed) were identified and transformed to soft data for BME modeling in this study.

3.3 Soft data generation based on outliers

Figure 5 shows the histogram of the 24 identified outliers (global and spatial outliers). These outliers were automatically divided into nine groups; most of them lie in the range of 150–300 mg/kg. The highest Zn concentration (529.71 mg/kg) was isolated from that of other outliers. A Gaussian distribution with a mean and SD parameter equal to that of the Zn concentration of the outliers, which was 224.73 and 87.23 mg/kg, respectively, was inadequate to represent the outliers. Therefore, a discrete PDF was used as soft data for BME modeling in this study, i.e., the sum of the area of bins in the histogram is equal to one.

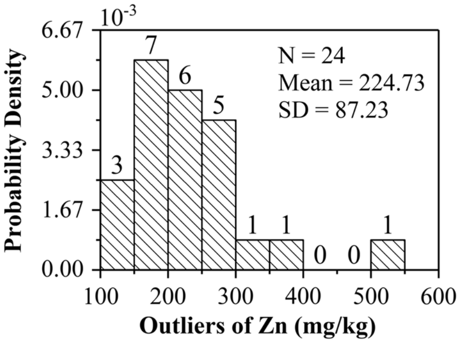


Fig. 5 Discrete probability soft data based on the identified outliers (N = 24). The number of outliers in each interval is displayed on the corresponding bin. The mean value and SD of the outliers are also shown

3.4 Comparisons of OK and BME for spatial prediction of soil Zn concentrations

After the outliers were identified and excluded, the robust variogram for OK and covariance model for BME can be fitted (Fig. S1) to perform the model validation and spatial prediction of Zn concentrations. For the OK method, the hard dataset (300 data points), along with the 24 outliers, was submitted to the robust variogram estimator. For the BME method, the hard and soft data (Fig. 5) were modeled simultaneously to obtain the spatial estimates of Zn concentrations at validation sites. The performances of OK and BME with the existence of outliers can then be compared on the basis of the creation described in Sect. 0.

Table 1 presents the cross-validation criteria, namely, ME and MSE. The ME value of the OK method was positive (5.64 mg/kg), whereas that of the BME method was negative (− 3.24 mg/kg). This result implied that OK was likely to overestimate the Zn concentrations (the positive errors were more than the negative errors), whereas the BME method would underestimate the Zn contents. However, the magnitude of the ME values was relatively small compared with the sampling Zn concentration, indicating that OK and BME were unbiased.

Nevertheless, the BME method produced a slightly higher MSE (569.97 mg^2/kg^2) than the OK method (554.47 mg^2/kg^2). The MSE components, which are listed in Table 1 for their values and proportion to MSE values, can provide additional information about the difference between the estimated and observed values. BME produced a larger SDSD than OK (306.35 mg^2/kg^2 for BME and 168.35 mg^2/kg^2 for OK), and this component (SDSD) contributed most to the MSE of BME (53.75%), which indicated that BME failed to estimate the magnitude of fluctuation in the measured soil Zn at validation sites. This result was expected because in OK, the outlier values were directly used in the prediction process, and the weighted estimates of OK would be elevated if the outliers were searched and counted. In BME, the outliers were transformed into discrete soft data (Fig. 5) with a mean of 224.73 mg/kg, which differs considerably with the largest observed Zn content (529.71 mg/kg), and an SD of 87.23 mg/kg, which is relative large to its mean value. Thus, the mean estimator adopted in this study (Eq. (13)) would correspond to the statistical characteristic of soft data and provide a narrower range of predicted values than OK at the validation sites (minimum and maximum predicted values of OK and BME in Table 1 and Fig. S2). The LCS is the component that contributed the most to the MSE of OK (63.90%), and its contribution was intermediate for BME (44.41%). This result suggested that OK failed to estimate the degree of fluctuation in the observed Zn concentrations.

Figure 6a, b show the results of Zn concentration predicted on the basis of OK and BME. The general trends of the spatial distribution of Zn generated using the two techniques were similar, indicating that the Zn contents were relatively high along the Yangtze River and decreased gradually with the distance from the river (the city center is along the river). However, the OK prediction is represented by a continuous and smooth surface mainly due to its smooth effect. On the contrary, some isolated circles could be found in the BME predictions (Fig. 6b), which may be the natural consequence of the substantial spatial heterogeneity of the Zn concentrations in the study area. To test the spatial heterogeneity, the q-statistic of the GeoDetector approach (Shi et al. 2015; Wang et al. 2016) was applied to the BME predictions with strata partitioned by land use types of Wuhan City (Fig. S3), which is an important anthropogenic factor affecting the spatial distribution of soil heavy metals. Results indicated the significant ($q = 0.15, p < 0.001$) effect of land use types on the spatial heterogeneity of Zn concentrations.

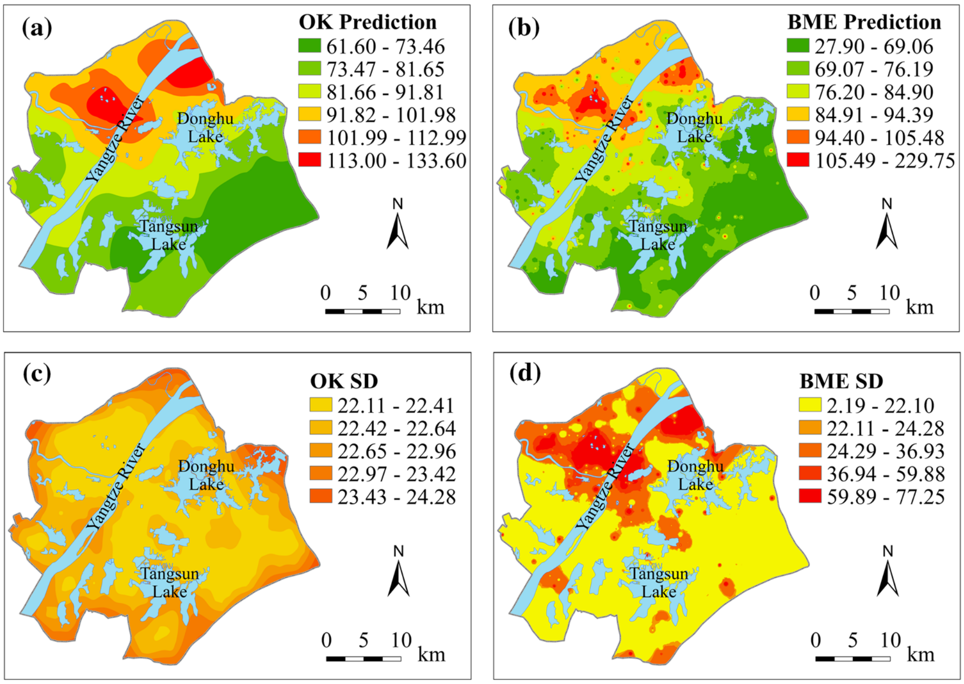
A comparison of Fig. 6a, b shows that BME had a considerably wider range of predicted values than OK, but it is not in conflict with the results of model cross-validation. On the one hand, most of the BME predictions lie in the range of 50–120 mg/kg, whereas that of OK is 60–130 mg/kg (Fig. S4). On the other hand, BME would honor the observed values as much as possible, whereas OK would probably generate smoothed results (the maximum value of BME prediction is 229.75 mg/kg, which is considerably larger than that of OK, 133.60 mg/kg).

The prediction error SD offers a method for examining the level of uncertainty in the resulting analyses. Figure 6c, d depict the spatial distribution of the prediction error SD generated by OK and BME. Figure 6c indicates that the prediction error SD of OK (OK SD) depicted a border effect (i.e., OK SD values were larger in the border region than in the central region), which was likely because the spatial extrapolation would occur as no data points exist near the border for OK. For the prediction error SD of

Table 1 Quantitative criteria for comparing the performances of OK and BME at validation sites

Method	Min (mg/kg)	Max	ME	MSE	SB	SDSD	LCS	SB (%)	SDSD	LCS
OK	63.84	123.58	5.64	554.47	31.84	168.35	354.28	5.74	30.36	63.90
BME	57.61	102.21	− 3.24	569.97	10.50	306.35	253.12	1.84	53.75	44.41

Fig. 6 Maps of prediction results: **a**, **b** predicted Zn concentrations by OK and BME and **c**, **d** prediction error SD by OK and BME across the study area



BME (BME SD) in Fig. 6d, no border effect was found. Instead, a large BME SD was likely to occur in the region where Zn predictions were high. This result is in accordance with the previous study that the BME SD depends on the specific set of data values considered, whereas OK SD only depends on the data configuration (Christakos and Serre 2000).

Considering the magnitude of the prediction error SD of OK and BME (Fig. 6c, d), the BME SD was smaller than the minimum value of OK SD (22.11 mg/kg) in a large area, as indicated by the light color in Fig. 6d. However, some regions had a BME SD larger than the maximum OK SD (24.28 mg/kg), especially the two zones located along the Yangtze River (dark red zones in Fig. 6d). The explanation for these results is as follows. On the one hand, as the OK method aims to minimize the mean squared estimation error (Olea 2006; Oliver and Webster 2014), its prediction SD should be maintained at a relatively low level. On the other hand, the uncertainty distribution of BME prediction is considerably affected by the quality of soft data, and increased level of uncertainty would occur when soft data with a large range were used (Christakos et al. 2004; Christakos and Serre 2000). In this study, the discrete soft data were characterized by a large SD (Fig. 5). Figures 4 and 6d indicate that more outliers were located in the two red zones in Fig. 6d than in other areas. Thus, soft data converted from these outliers would be counted in the BME prediction in the two red zones, thereby elevating the BME prediction SD.

Nevertheless, the advantage of the BME method can still be found by observing the PDF at selected validation sites (Fig. 7). Figure 7a, b plot the PDF of OK and BME predictions at the validation sites, where the minimum and maximum BME SD occurred, respectively. The PDF peaks for OK and BME predictions did not differ considerably, but the PDF range of BME prediction was considerably narrow (Fig. 7a). The PDF of BME predictions was considerably close to the observed Zn contents (PDF peak and vertical solid line) although the PDF range of OK prediction was narrow (Fig. 7b). From Fig. 7c, the BME prediction matched the observed Zn contents when the BME prediction error was small. In Fig. 7d, the probability that BME prediction covered the observed Zn contents was considerably higher than that of OK prediction, although the Zn concentration predicted using the BME method was further away from that using the OK method. This result indicates that the BME method can provide considerable informative results through its posterior PDF by incorporating sufficient information provided by soft data. The outliers are probably generated by the mechanisms different from those of the other data, and the robust variogram $\hat{\gamma}_{CH}(h)$ successfully utilizes this hypothesis, such that the outliers are modeled as second processes (Lark 2000). However, this mathematical solution to lower the effect of outliers may only reduce the contribution of useful information by outliers. On the contrary, the BME method directly acquires the highly uncertain information carried by outliers through soft data and therefore improves the spatial prediction.

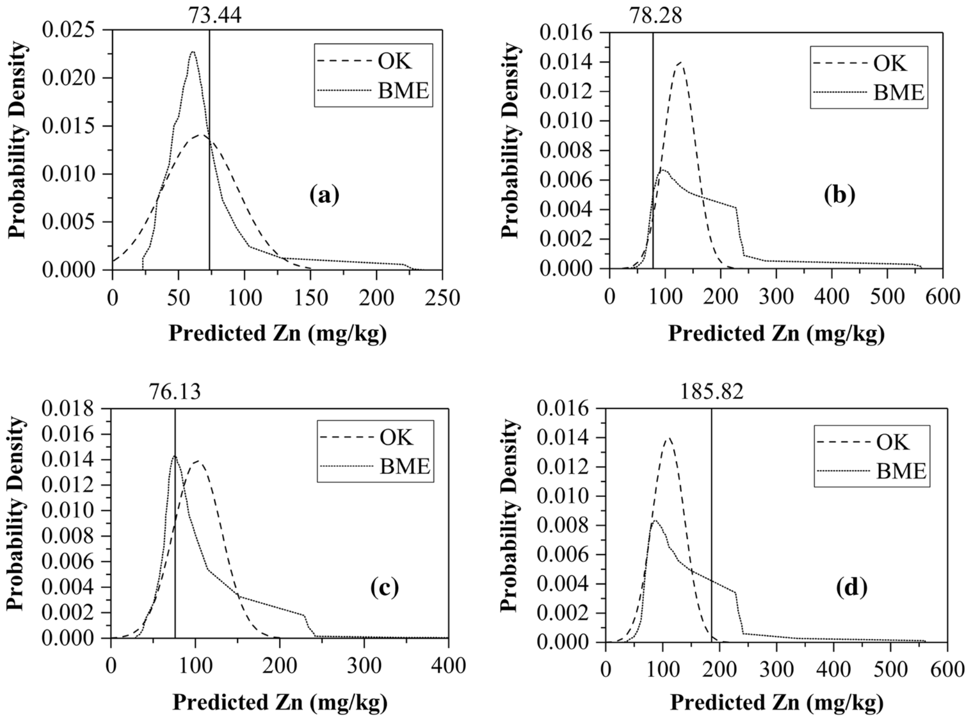


Fig. 7 PDF of OK prediction and the corresponding BME posterior PDF at the validation sites, where BME had the **a**, **b** minimum and maximum prediction standard error variances, respectively, and **c**, **d** minimum and maximum prediction errors, respectively. The number on the top of the solid vertical line represents the observed Zn concentrations

4 Conclusions

This study investigates the feasibility of the BME method in the spatial prediction of soil heavy metals with the existence of outliers. Estimation results indicated that the BME method, which incorporates outliers as soft data, could produce more accurate prediction results than the OK method. In addition, the PDF of BME predictions was further informative and matched the observed Zn concentrations well. However, in this case study, BME failed to estimate the magnitude of fluctuation in the measured soil Zn at validation sites, and its prediction SD was larger than that of OK method in some of the regions of the study area. This result was mainly caused by the relatively high uncertainty of soft data. An improved performance of BME could be expected if further appropriate forms of soft data were supported by the software or if additional high-quality soft data were constructed. Therefore, the BME method is promising in managing environmental data where outliers are relatively often encountered.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant Nos. 41701239 and 41671217) and the Fundamental Research Funds for the Central Universities (No. 2452016190).

References

Benhaddya M, Hadjel M (2014) Spatial distribution and contamination assessment of heavy metals in surface soils of Hassi Messaoud, Algeria. Environ Earth Sci 71(3):1473–1486

China National Environmental Monitoring Centre (1990) The background values of soil elements in China. China Environmental Science Press, Beijing **(in Chinese)**

Christakos G (1990) A Bayesian/maximum-entropy view to the spatial estimation problem. Math Geol 22(7):763–777

Christakos G (2000) Modern spatiotemporal geostatistics. Oxford University Press, New York

Christakos G, Li XY (1998) Bayesian maximum entropy analysis and mapping: a farewell to kriging estimators? Math Geol 30(4):435–462

Christakos G, Serre ML (2000) BME analysis of spatiotemporal particulate matter distributions in North Carolina. Atmos Environ 34(20):3393–3406

Christakos G, Kolovos A, Serre ML, Vukovich F (2004) Total ozone mapping by integrating databases from remote sensing instruments and empirical models. IEEE Trans Geosci Remote Sens 42(5):991–1008

Cressie N, Hawkins DM (1980) Robust estimation of the variogram: I. J Int Assoc Math Geol 12(2):115–125

Douaik A, Van Meirvenne M, Toth T (2005) Soil salinity mapping using spatio-temporal kriging and Bayesian maximum entropy with interval soft data. Geoderma 128(3–4):234–248

Gao SG, Zhu ZL, Liu SM, Jin R, Yang GC, Tan L (2014) Estimating the spatial distribution of soil moisture based on Bayesian maximum entropy method with auxiliary data from remote sensing. Int J Appl Earth Obs 32:54–66

Guo GH, Wu FC, Xie FZ, Zhang RQ (2012) Spatial distribution and pollution assessment of heavy metals in urban soils from southwest China. J Environ Sci China 24(3):410–418

Helmreich B, Hilliges R, Schriewer A, Horn H (2010) Runoff pollutants of a highly trafficked urban road—correlation analysis and seasonal influences. Chemosphere 80(9):991–997

Lark RM (2000) A comparison of some robust estimators of the variogram for use in soil survey. Eur J Soil Sci 51(1):137–157

Matheron G (1963) Principles of geostatistics. Econ Geol 58(8):1246–1266

McGrath D, Zhang C (2003) Spatial distribution of soil organic carbon concentrations in grassland of Ireland. Appl Geochem 18(10):1629–1639

Meklit T, Meirvenne MV, Verstraete S, Bonroy J, Tack F (2009) Combining marginal and spatial outliers identification to optimize the mapping of the regional geochemical baseline concentration of soil heavy metals. Geoderma 148(3):413–420

Meza-Figueroa D, De la O-Villanueva M, De la Parra ML (2007) Heavy metal distribution in dust from elementary schools in Hermosillo, Sonora, México. Atmos Environ 41(2):276–288

Ministry of Ecology and Environment of the People's Republic of China (2018) Soil environmental quality risk control standard for soil contamination of agricultural land **(in Chinese)**

Olea RA (2006) A six-step practical approach to semivariogram modeling. Stoch Environ Res Risk A 20(5):307–318

Oliver MA, Webster R (2014) A tutorial guide to geostatistics: computing and modelling variograms and kriging. CATENA 113:56–69

Puangthongthub S, Wangwongwatana S, Kamens RM, Serre ML (2007) Modeling the space/time distribution of particulate matter in Thailand and optimizing its monitoring network. Atmos Environ 41(36):7788–7805

Reyes JM, Serre ML (2014) An LUR/BME framework to estimate PM2.5 explained by on road mobile and stationary sources. Environ Sci Technol 48(3):1736–1744

Savelieva E, Demyanov V, Kanevski M, Serre M, Christakos G (2005) BME-based uncertainty assessment of the chernobyl fallout. Geoderma 128(3–4):312–324

Shi TT, Yang XM, Christakos G, Wang JF, Liu L (2015) Spatiotemporal interpolation of rainfall by combining BME theory and satellite rainfall estimates. Atmosphere 6(9):1307–1326

Wang JF, Zhang TL, Fu BJ (2016) A measure of spatial stratified heterogeneity. Ecol Indic 67:250–256

Webster R, Oliver MA (2007) Geostatistics for environmental scientists. Wiley, Hoboken

Wuhan Bureau of Statistics (2013) The Wuhan Statistical Bulletin of National Economic and Social Development **(in Chinese)**

Xu CD, Wang JF, Hu MG, Li QX (2014) Estimation of uncertainty in temperature observations made at meteorological stations using a probabilistic spatiotemporal approach. J Appl Meteorol Climatol 53(6):1538–1546

Yu H-L, Kolovos A, Christakos G, Chen J-C, Warmerdam S, Dev B (2007) Interactive spatiotemporal modelling of health systems: the SEKS-GUI framework. Stoch Environ Res Risk A 21(5):555–572

Zhang CS, Tang Y, Luo L, Xu WL (2009) Outlier identification and visualization for Pb concentrations in urban soils and its implications for identification of potential contaminated land. Environ Pollut 157(11):3083–3090

Zhang CT, Yang Y, Li WD, Zhang CR, Zhang RX, Mei Y, Liao XS, Liu YY (2015) Spatial distribution and ecological risk assessment of trace metals in urban soils in Wuhan, central China. Environ Monit Assess 187(9):1–16

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.