Chin. J. Geochem. (2015) 34(3):416–421
DOI 10.1007/s11631-015-0050-x

ORIGINAL ARTICLE

Environmental baseline evaluation of lead in shallow groundwater based on statistical and spatial outlier identification

Linhua Sun[1]

Received: 26 May 2014 / Revised: 23 June 2014 / Accepted: 30 June 2014 / Published online: 27 March 2015

Abstract A series of methods have been applied for evaluating the environmental baseline—especially on a regional scale—because of its importance for local environmental management. However, most of the methods have been statistical in nature, and spatial variability has not been considered simultaneously. In this study, the combined use of statistical and spatial methods has been carried out to analyse lead concentrations in the shallow groundwater in the urban area of Suzhou, northern Anhui Province, China. The environmental baseline of lead in the shallow groundwater has been evaluated to be 3.836–8.240 μg/l after removal of either statistical or spatial outliers. The results are similar to those obtained by model-based objective methods, and further demonstrate that the combination method is more reasonable for environmental baseline evaluation.

Keywords Environmental baseline · Shallow groundwater · Spatial cluster · Outlier identification · Lead

1 Introduction

Concentrations of elements in the natural environment can be affected by multiple processes, including natural (e.g. weathering) and anthropogenic (e.g. pollution) (Meklit et al. 2009). If the concentration exceeds a given reference value, known as environmental background (Hawkes and Webb 1962), the area is considered to be contaminated. However, the reference value is at a global (e.g. WHO 2008) or national (e.g. CEPA 1990) scale, and therefore might be meaningless when applied to a regional scale since the natural levels of elements vary significantly from area to area.

However, as determined by previous studies, the environmental background, which is defined under the conditions of the natural environment without any anthropogenic contribution, is difficult to obtain because almost the entire world has been affected by human activities. As an alternative, the concept of environmental baseline has been put forward (Salminen and Tarvainen 1997; Reimann and Garrett 2005), and a large number of studies have been carried out for evaluating baseline values, especially at a regional scale.

Environmental baseline can be quantified in several ways with different methods (Reimann and Garrett 2005), and the most important one is statistical. For example, some statistical methods assume the normality or log-normality of concentration distribution (Reimann and Garrett 2005; Galuszka 2007), and cumulative probability plots, as well as Q–Q plots, have been applied for data analysis. Moreover, some previous studies revealed that it is more realistic to view geochemical baseline as a range of values rather than an absolute value because it changes both regionally with the basic geology and locally with the type and genesis of overburden. In prior studies, model-based objective methods (iterative 2σ technique and the calculated distribution function) have been applied (Nakic et al. 2007; Sun 2013; Urresti-Estala et al. 2013).

Suzhou is a city in the northern Anhui Province, China, that is dominated by agriculture and coal production. Groundwater is important for industrial, agricultural, and domestic use in the area because of the lack of surface water: the annual rainfall is only 774–895 mm and

✉ Linhua Sun
sunlinh@126.com

[1] School of Resources and Civil Engineering, Suzhou University, 71 Bianhe Road, Suzhou 234000, People's Republic of China

concentrated in the period from May to September. Even in the urban area, about 30 % of residences use groundwater pumped from shallow wells (<30 m depth). Lead in water has long been a concern of environmental scientists (Goldberg 1974; Darling and Thomas 2003), and lead in drinking water can cause a variety of adverse health effects (e.g. damage to the brain and nervous system, increased blood pressure). Although groundwater is an important source of water for drinking and other uses in Suzhou, the geochemical baseline has not yet been determined.

Therefore, in this study, lead concentrations in shallow groundwater in the urban area of Suzhou have been measured and the data have been analyzed by statistical and spatial methods. The goals of the study include: (1) identifying the outlier samples and (2) evaluating the environmental baseline.

2 Materials and methods

2.1 Sampling and analysis

In this study, a total of 62 groundwater samples were collected from shallow wells (<30 m depth) in the urban area of the city (Fig. 1) between September and October, 2013.

The samples were first filtered using 0.45 μm pore-size membranes into 2.0-l polyethylene bottles that had been cleaned in the laboratory, and then immediately acidified to $pH < 2$ with HNO_3 to prevent the precipitation and/or adsorption of elements by the bottle. The samples were sent for analysis within 24 h of collection. Analysis was carried out at the Engineering and Technology Research Center of Coal Exploration in Anhui Province, China. An atomic absorption spectrometer was used for analyzing the concentration of lead. A calibration curve was obtained using a series of different concentrations of lead standard; the coefficient of the curve is 0.99.

2.2 Data analyses

All of the lead concentrations were first analyzed by the software Mystat (version 12), and the minimum, maximum, mean, standard deviation, coefficient of variation, and the p value of Anderson–darling normality test obtained. Then, the software Surfer (version 11) was applied to produce a contour map of lead concentrations, and kriging was chosen for the gridding method. Finally, the software GeoDa (version 1.4.6) was applied for spatial analysis. The box plot and map with Hinge = 1.5 was applied for statistical outlier identification. With this procedure, the lower and upper outliers were identified.

Next, spatial cluster analysis, which is named Univariate Local Moran's I in the GeoDa software, was applied to the dataset, and five categories (including not significant, high–high, low–low, low–high, and high–low) were obtained. During this procedure, samples in the high–high cluster were identified as hotspots, whereas samples in high–low and low–high clusters were considered to be outliers. In comparison with other hotspot identification methods (e.g. Getis's G index, spatial scan statistics, and Tango's C index) (Getis and Ord 1992; Ishioka et al. 2007; Tango 1995), the Moran's I index examines the individual locations, enabling hotspots to be identified based on comparison with neighboring samples. After removing the outliers obtained by either box plot or spatial analysis, the mean $\pm 2\sigma$ (Nakic et al. 2007) of the rest of the samples was then considered to be the baseline value. During the spatial analysis, rook contiguity was chosen for weight calculation.

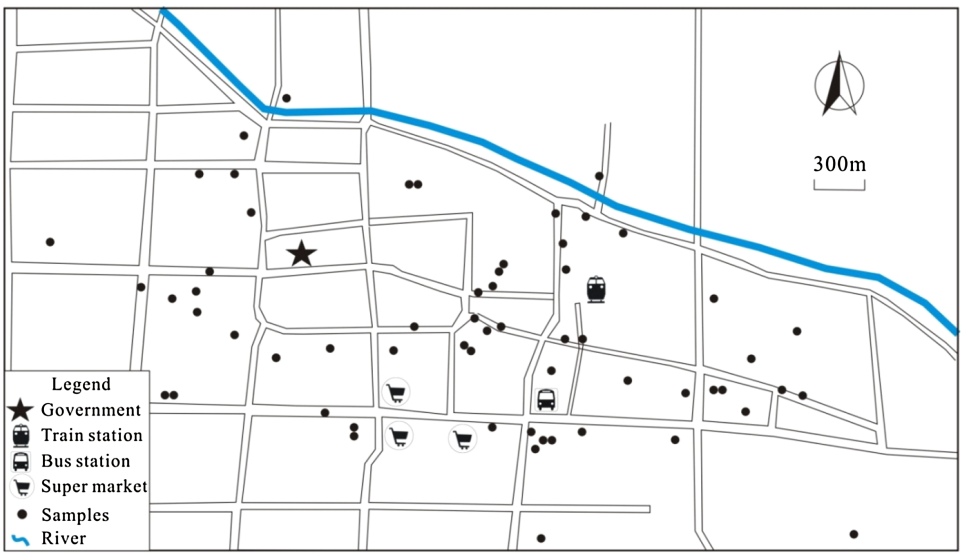

Fig. 1 Sample locations in the study area

3 Results and discussion

3.1 Descriptive statistics

The descriptive statistics of the lead concentrations (μg/l) are listed in Table 1. As can be seen from the table, the samples in this study have lead concentrations ranging from 4.161 to 11.526 μg/l. Their mean and median values are 6.583 and 6.299 μg/l, respectively. According to the quality standards for groundwater in China (μg/l, GB/T 14848-9), the samples in this study can be subdivided into three categories: eight samples in class I (≤5 μg/l), 51 samples in class II (≤10 μg/l), and three samples in class III (≤50 μg/l). Such a result indicates that all samples can be used for drinking, irrigation, and industry directly (only considering about their lead concentrations). Moreover, the spatial distribution of the lead concentrations in the shallow groundwater in this study has a low-to-moderate coefficient of variation (0.253), implying that the shallow groundwater system in the area has not been dramatically affected by human activities.

However, the p value of the Anderson–Darling normality test is less than 0.05, meaning that the lead concentrations of the samples in this study cannot pass the normality test, which indicates the possibility of an anthropogenic contribution (Reimann and Garrett 2005). This possibility is also demonstrated by the contour plots of the lead concentrations in the area (Fig. 2), in which two centers with high lead concentrations can be identified: one is located from the east to the center of the map, and another is located in the west. Alternatively, these areas with high lead concentration might be the results of geological heterogeneity.

3.2 Outlier identification by box plot

Previous studies reveal that the geochemical baseline and the pollution data are different in both their statistical distribution and spatial behavior (Meklit et al. 2009). Therefore, the box plot, a convenient way of graphically depicting groups of numerical data through their quartiles, has been employed for identifying the statistical outliers. The method has long been used for outlier identification during environmental background or baseline studies (Reimann and Garrett 2005). It is a statistical method in which the lower and upper inner fences are calculated (see functions 1 and 2, respectively) (Frigge et al. 1989), and the samples with higher or lower concentrations relative to the fences are considered to be outliers.

$$\text{Function 1}: 25\,\%\ \text{percentile} - 1.5 \times (75\,\%\ \text{percentile} - 25\,\%\ \text{percentile}).$$

$$\text{Function 2}: 75\,\%\ \text{percentile} + 1.5 \times (75\,\%\ \text{percentile} - 25\,\%\ \text{percentile}).$$

Based on these functions, the lower and upper inner fences of the lead concentrations in this study were calculated to be 3.092 and 9.744 μg/l, respectively, and four samples (samples 23, 49, 50, and 60) with lead concentrations higher than 9.744 μg/l have been identified as outliers; their locations are shown in Fig. 3 as a box map. As can be seen from the figure, these four samples are located in areas with high lead concentrations (Fig. 2), which indicates that these areas might have been affected by human activities.

3.3 Outlier identification by spatial cluster

Similar to the statistical outliers, samples with unusual values relative to their neighborhood are also considered to be outliers, and are known as spatial outliers (Lark 2002). To identify spatial outliers, a series of methods have been applied. For instance, variograms were used to model the spatial autocorrelation with a cross-validation procedure of ordinary kriging, an estimated value was generated for every measurement, and then the standardized estimation error was used for identifying spatial outliers (Laslett and Mabratney 1990). Moreover, Moran's I is a commonly used indicator of spatial autocorrelation and two types of Moran's I have been reported for different destinations: the global Moran's I was used to study the overall spatial

Table 1 Summary statistics of the complete dataset and of the resulting dataset after outlier removal by statistical and spatial methods (unit for concentration: μg/l)

	Whole data	Box plot	Spatial outlier	Combination
N of cases	62	58	52	50
N of outliers	0	4	10	12
Minimum	4.161	4.161	4.161	4.161
Maximum	11.526	9.388	10.813	9.388
Median	6.299	6.180	6.061	5.943
Mean	6.583	6.287	6.203	6.038
Standard deviation	1.667	1.251	1.368	1.101
Coefficient of variation	0.253	0.199	0.221	0.182
p value	<0.01	0.098	<0.01	>0.15

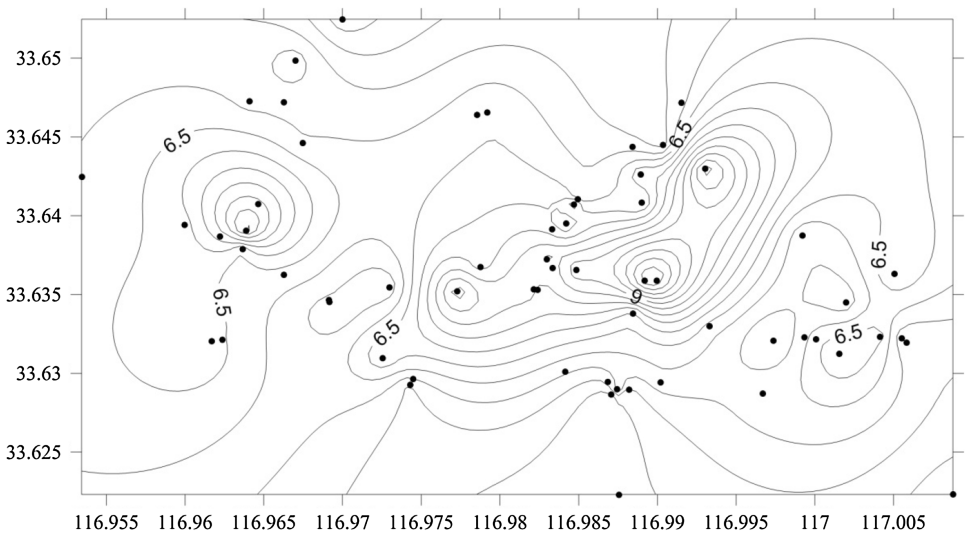

Fig. 2 Contour map of lead concentrations

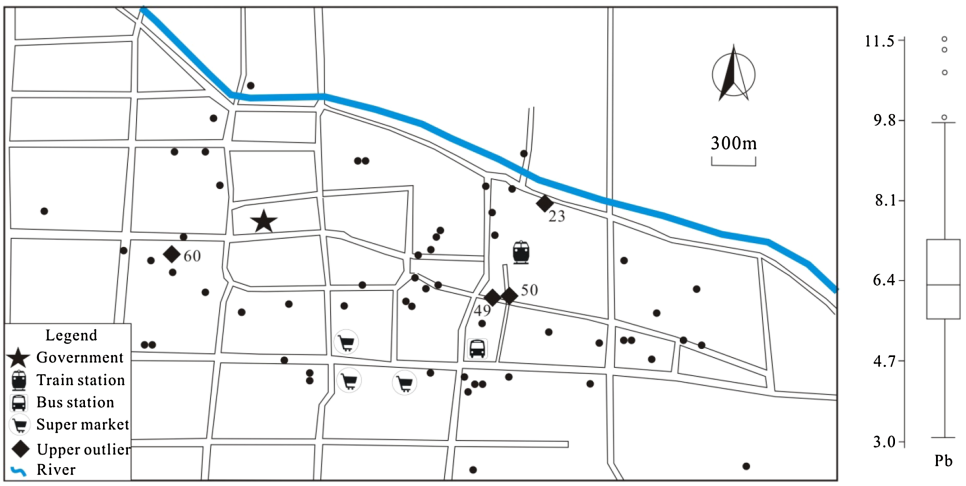

Fig. 3 Outlier distribution based on box plot (*box map*)

autocorrelation, whereas LISA (local indicators of spatial association) were applied to identify the degree of spatial autocorrelation in each specific location by using Local Moran's I (Anselin 1995). It can also be used for identifying the existence of local spatial clusters by generating cluster maps (Zhang and McGrath 2004; Harries 2006).

In this study, all of the samples have been classified as not significant (46 samples) or significant (16 samples). Among the significant samples, eleven and five samples have a significance level of p value equal to 0.05 and 0.01, respectively. Moreover, the significant samples are classified into four secondary categories: high–high (seven samples), low–low (six samples), low–high (one sample), and high–low (two samples).

Spatial distribution of heavy metal concentrations is useful for identifying "hotspots" (Li et al. 2014). According to previous studies (Zhang et al. 2008), either high–high or low–low samples can be grouped in spatial clusters, whereas high–low and low–high samples are considered to be spatial outliers. As can be seen from Fig. 4, two "hotspots" can be identified: one is located in the center of the map and another is located in the west of the map. This is similar to the results obtained by contour map (Fig. 2) and might be an indication of special human activities. For instance, the east-center hotspot is located in the area of an old, densely populated quarter, and the train and bus stations are located in the area. Moreover, some small workshops with steel, iron, and lead products are also located in the area (Fig. 4). In summary, the high–high, high–low, and low–high samples—ten samples in total (samples 10, 24, 26, 32, 38, 39, 41, 49, 50, and 51)—are considered to be spatial outliers.

3.4 Environmental baseline evaluation

After box plot identification, the samples with extreme high values were removed. However, this procedure does not take the spatial variability into account, and it is incapable of

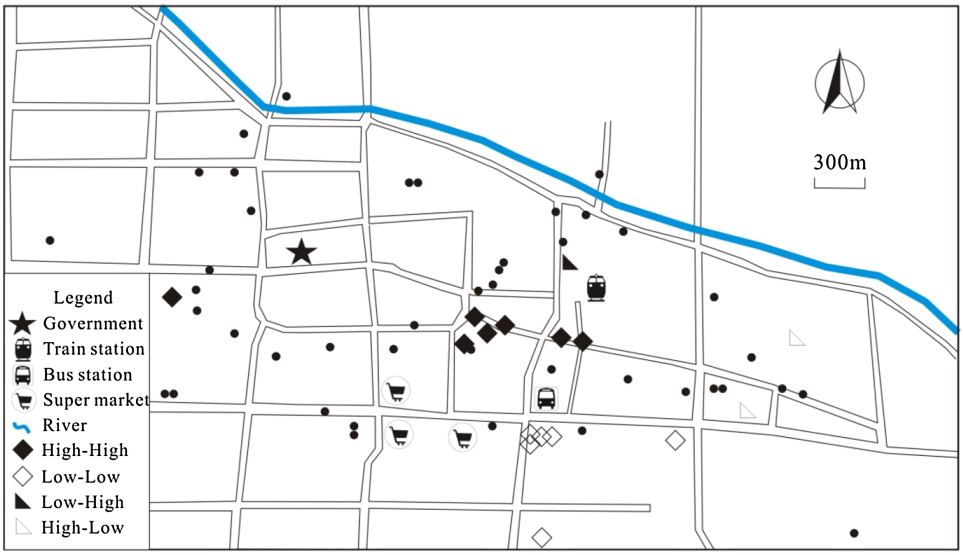

Fig. 4 Outlier distribution based on spatial analysis

defining the unique environmental baseline on a local scale (Meklit et al. 2009). Therefore, the combined use of the two methods (box plot and spatial cluster) can produce more reliable information, as it can remove both the statistical and spatial outliers. In consideration of this, a total of twelve samples are classified as outliers, and the remaining fifty samples are considered to be environmental baseline samples. The summary statistics of the samples with different outlier identifying methods are shown in Table 1.

As can be seen from the table, after outlier removal by box map, 58 samples remain, and their mean concentration is 6.287 μg/l. However, after spatial outlier removal, 52 samples remain, and their mean concentration is 6.203 μg/l. The remaining samples after statistical outlier removal can pass the normality test with p value = 0.098, whereas the remaining samples after spatial outlier removal cannot pass the normality test. This is due to the differences between these two methods; the former assumes a normal distribution of the baseline data, whereas the latter considers only the concentration variations among neighborhoods.

As for the combined method, the mean concentration is 6.038 μg/l with a standard deviation of 1.101 μg/l, and the environmental baseline is therefore established to be 3.836–8.240 μg/l. This is similar to the results obtained by using model-based objective methods (iterative 2σ technique and the calculated distribution function) (Nakic et al. 2007; Urresti-Estala et al. 2013), by which the baseline values were determined to be 4.1–7.8 and 4.2–8.4 μg/l, respectively.

4 Conclusions

Based on the combined use of statistical and spatial analyses of lead concentrations in shallow groundwater collected from the urban area of Suzhou, northern Anhui Province, China, the following conclusions have been made:

(1) The lead concentrations in the groundwater samples are low, and they can be used for drinking, irrigation, and industry according to the Chinese and WHO standards;
(2) Four outlier samples with the highest lead concentrations have been identified by box plot and map, whereas ten samples have been identified as outliers by spatial analyses;
(3) Two hotpots, which are located in the center and west of the map, have been identified and might be an indication of special human activities;
(4) The environmental baseline based on the dataset of samples after removal of statistical and spatial outliers is estimated to be 3.836–8.240 μg/l, and is similar to the results obtained by model-based objective methods.

Acknowledgments This work was financially supported by the National Natural Science Foundation of China (41302274 and 41173106), and the National College Students Innovation and Entrepreneurship Training Program of China (201210379026).

References

Anselin L (1995) Local indicators of spatial association-LISA. Geogr Anal 27:93–115

CEPA (Chinese Environmental Protection Administration) (1990) Elemental background values of soils in China. Environmental Science Press, Beijing

Darling CTR, Thomas VG (2003) The distribution of outdoor shooting ranges in Ontario and the potential for lead pollution of soil and water. Sci Total Environ 313(1):235–243

Frigge M, Hoaglin DC, Iglewicz B (1989) Some implementations of the Boxplot. Am Stat 43(1):50–54

Gałuszka A (2007) A review of geochemical background concepts and an example using data from Poland. Environ Geol 52(5):861–870

Getis A, Ord JK (1992) The analysis of spatial association by use of distance statistics. Geogr Anal 24:189–206

Goldberg A (1974) Drinking water as a source of lead pollution. Environ Health Perspect 7:103–105

Harries K (2006) Extreme spatial variation in crime density in Baltimore County, MD. Geoforum 37:995–1017

Hawkes HE, Webb JS (1962) Geochemistry in mineral exploration. Harper and Row, New York

Ishioka F, Kurihara K, Suito H, Horikawa Y, Ono Y (2007) Detection of hotspots for three-dimensional spatial data and its application to environmental pollution data. J Environ Sci Sustain Soc 1:15–24

Lark RM (2002) Modeling complex soil properties as contaminated regionalized variables. Geoderma 106:173–190

Laslett GM, Mabratney AB (1990) Further comparison of spatial methods for predicting soil-pH. Soil Sci Soc Am J 54:1553–1558

Li W, Xu B, Song Q, Liu X, Xu J, Brookes PC (2014) The identification of 'hotspots' of heavy metal pollution in soil-rice systems at a regional scale in eastern China. Sci Total Environ 472:407–420

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

Nakic Z, Posavec K, Bacani A (2007) A visual basic spreadsheet macro for geochemical background analysis. Ground Water 45(5):642–647

Reimann C, Garrett RG (2005) Geochemical background—concept and reality. Sci Total Environ 350(1):12–27

Salminen R, Tarvainen T (1997) The problem of defining geochemical baselines. A case study of selected elements and geological materials in Finland. J Geochem Explor 60(1):91–98

Sun L, Gui H, Peng W, Lin M (2013) Heavy metals in deep seated groundwater in northern Anhui Province, China: quality and background. Nat Environ Pollut Technol 12(3):533–536

Tango T (1995) A class of tests for detecting 'general' and 'focused' clustering of rare diseases. Stat Med 14:2323–2334

Urresti-Estala B, Carrasco-Cantos F, Vadillo-Pérez I, Jiménez-Gavilán P (2013) Determination of background levels on water quality of groundwater bodies: a methodological proposal applied to a Mediterranean River basin (Guadalhorce River, Malaga, southern Spain). J Environ Manag 117:121–130

WHO (World Health Organization) (2008) Guidelines for drinking-water quality (3rd edition). World Health Organization, Geneva

Zhang C, McGrath D (2004) Geostatistical and GIS analyses on soil organic carbon concentrations in grassland of southeastern Ireland from two different periods. Geoderma 119:261–275

Zhang C, Luo L, Xu W, Ledwith V (2008) Use of local Moran's I and GIS to identify pollution hotspots of Pb in urban soils of Galway, Ireland. Sci Total Environ 398:212–221