Nature Environment and Pollution Technology
An International Quarterly Scientific Journal

ISSN: 0972-6268	Vol. 14	No. 2	pp. 373-378	2015

Original Research Paper

Application of Statistical and Spatial Outlier Identification for Evaluating the Environmental Baseline of Iron in Shallow Groundwater

Linhua Sun† and Herong Gui
School of Resource and Civil Engineering, Suzhou University, 234000, Suzhou, China
†Corresponding author: Linhua Sun

Nat. Env. & Poll. Tech.
Website: www.neptjournal.com
Received: 25-7-2014
Accepted: 20-8-2014

Key Words:
Baseline of iron
Shallow groundwater
Spatial analysis
Outlier identification

ABSTRACT

Environmental baseline is essential for local environmental management, and a series of methods have been carried out for solving this issue. In this study, sixty-two shallow groundwater samples from the urban area in Suzhou, northern Anhui Province, China have been collected and analysed for their iron concentrations, and then processed by either statistical (box plot) or spatial analyses (spatial clustering) for outlier identification. The results indicate that four and five samples have been identified as outliers by box plot and spatial analysis, respectively, and the rest of the samples (fifty-four) have been set as environmental baseline samples. Their mean ± 2σ concentration is then set as environmental baseline (0-104.720 µg/L). The study demonstrated that spatial analysis is useful for assisting the outlier identification during evaluating the environmental baseline relative to statistical methods.

INTRODUCTION

The environmental background is important for local environmental management, because it has been set as the criterion for evaluating the pollution degrees in areas. Nowadays, most of the environmental backgrounds, such as the suggestions of World Health Organization (WHO 2008), are determined in a global or national scale (e.g. CEPA 1990), and they have been widely used for the studies of environmental pollution or management (Sun et al. 2012, Wongsasuluk et al. 2014). However, although these criterions are good choice for a global or national comparison, it is not suitable for the environmental studies or management in a regional scale, because the local environmental background, which is determined by local parent materials, can vary significantly from area to area. And therefore, a simple application of global or national environmental background can lead to the wrong understanding about the pollution degrees.

Nowadays, a large number of studies (Molinari et al. 2012) related to the determination of local environmental background have been processed. However, most of the previous studies found that the background, which means the natural condition without any anthropogenic contribution, is hard to obtain because of the long history of human activities. And therefore, the concept of environmental baseline, which was first proposed by Salminen & Tarvainen (1997), has been put forward, and a large number of studies have been carried out for evaluating the baseline value, especially in a regional scale (Reimann & Garrett 2005).

A series of techniques, including statistical and non-statistical, have been carried out for evaluating the environmental baseline (Reimann et al. 2005). Taking for instance, the cumulative probability plot and Q-Q plot, which assume the normality or log-normality of concentration distribution, have long been used to determine the environmental baseline (Reimann & Garrett 2005, Galuszka 2007). Moreover, some other methods, such as the model based objective methods (including iterative 2σ technique and the calculated distribution function) (Nakic et al. 2007, Sun et al. 2013, Urresti-Estala et al. 2013) have also been applied, and they 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. These studies demonstrated that if we want to get reliable information about the environmental baseline, the outliers should be first considered.

In this study, the geo-statistical and spatial analysis, together with the aid from geographic information system (GIS), have been applied for the iron concentrations in shallow groundwater in the urban area of Suzhou, northern Anhui Province, China. The goals of the study include: (1)

identifying the concentration outliers (samples) and (2) establishing the environmental baseline of iron in the shallow groundwater.

MATERIALS AND METHODS

Sampling and analysis: As an agriculture and coal production dominated city, Suzhou is located in the northern Anhui Province, China. The annual rainfall in the city is only 774-895 mm and most of them are concentrated mainly in the period of May to September. And therefore, the total amount of surface water is limited and, the groundwater is important for the industrial, agricultural and domestic use. This phenomenon can be demonstrated by the well distribution in the urban area where about 30% of the urban residents use water by pumping from the shallow wells (less than 30 m) in the urban area. In this study, these shallow wells have been taken for research objects and, during the period between September and October 2013, a total of sixty-two shallow groundwater samples have been collected in the urban area of the city (Fig. 1).

All of the samples were collected following the standard procedures: firstly, they were filtered with 0.45 μm pore-size membranes before collecting into a 2.0 L polyethylene bottles that have been cleaned in the laboratory; secondly, they were immediately acidified to pH < 2 with HNO_3 for preventing the precipitation and/or adsorption of elements by the bottle. Finally, all the samples were immediately stored in a portable refrigerator and then sent for analysis within 24 hours. Analysis was carried out at the Engineering and Technology Research Center of Coal Exploration in Anhui Province, China. Atomic absorption spectrometer (AAS) was used for analysing the concentration of iron. Calibration curve was obtained using a series of different concentrations of lead standard and the coefficient of the curve was 0.99.

Data analyses: The procedures of data analyses were as follows: firstly, all the iron concentrations were processed by the software Mystat (version 12), and the minimum, maximum, mean, standard deviation, coefficient of variation, skewness and significant value of Anderson-Darling normality test have been obtained. Then, the software Surfer (version 11) has been applied for producing the contour map of iron concentrations, the gridding method is chosen for kriging. Finally, the software GeoDa (version 1.4.6) has been applied for spatial analysis: The box plot and map with Hinge = 1.5 (similar to the Box-plot in Reimann & Garrett 2005) has been applied for statistical outlier identification. With this procedure, the lower and upper outliers can be identified. And then, the spatial cluster analysis, which names Univariate Local Moran's I in the GeoDa software, has been applied for the dataset, and five categories (including not significant, high-high, low-low, low-high and high-low) can be obtained. The samples in high-high cluster were determined as hotspot samples, whereas samples in low-high and high-low clusters were selected as outliers. After removing the outliers obtained by either box map or spatial analysis, the mean ± 2σ (Nakic et al. 2007) of the rest of the samples was then considered to be baseline values. During the spatial analysis, rook contiguity was chosen for weight calculation. In comparison with other hotspot identification method (e.g. Getis's G index, spatial scan statistics and Tango's C index) (Getis & Ord 1992, Ishioka et al. 2007, Tango 1995), the Moran's I index examines the individual locations, enabling hotspots to be identified based on the comparison with the neighbouring samples.

RESULTS AND DISCUSSION

Descriptive statistics: The descriptive statistics of the iron concentrations (μg/L) are synthesized in Table 1. As can be seen from the table, the iron concentrations of the samples in this study have a broad range from 1.643 to 272.690 μg/L. Their mean and median values are 58.952 and 48.049 μg/L, respectively. Based on the quality standards for groundwater in China (μg/L, GB/T 14848-93), the samples in this study can be subdivided into three categories: most of the samples (fifty-three) are classified to be class I (≤ 100 μg/L), and eight samples are classified to be class II (≤ 200 μg/L), whereas only one sample is classified to be class III (≤ 300 μg/L). This result suggests that all of them can meet the requirement for drinking, irrigation and industry directly based on their iron concentrations. Even in comparison with the WHO standard (0.3 mg/L, WHO 2008), all the samples are suitable for drinking.

Moreover, the spatial distribution of the iron concentrations in the shallow groundwater in this study shows moderate-high coefficient of variation (0.801), implying that the shallow groundwater system in the area might have been affected by human activities, although the groundwater in it still has good quality. Alternatively, it might be the result of geological heterogeneity in the groundwater aquifer (e.g. wall rock variation or, different extents of water-rock interaction).

It can also be obtained from Table 1 that the p-value of Anderson-Darling normality test is less than 0.05, implying that the iron concentrations of the samples in this study cannot pass the normality test, which might be a suggestion of the existence of anthropogenic contribution (Reimann & Garrett 2005). This consideration is further supported by the contour plots of the iron concentrations in the area (Fig. 2), in which a series of centres with high iron concentrations can be identified, and most of them are presented in the cen-

tral of the map with a linear distribution.

Box plot-outlier identification: A series of studies revealed that the geochemical baseline and the pollution data are different in both their statistical and spatial distributions (Reimann & Garrett 2005, Meklit et al. 2009). For identifying the statistical differences, the box plot, a convenient way of graphically depicting groups of numerical data through their quartiles, has long been used for outlier selection during environmental background or baseline studies (Reimann & Garrett 2005). It is a statistical method by calculating the lower and upper inner fences (see in function 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.

Function 1: 25% percentile – 1.5 × (75% percentile - 25% percentile)

Function 2: 75% percentile + 1.5 × (75% percentile - 25% percentile)

In this study, all iron concentrations were firstly examined by box plot for identifying the statistical outliers (Fig. 3). Based on the functions listed above, the lower and upper inner fences of the iron concentrations in this study were calculated to be -30.805 and 136.755 μg/L, and only four samples (sample 27, 35, 56 and 57) with iron concentrations higher than 136.755 μg/L have been identified as outliers. For a more clear presentation, their locations are shown in Fig. 3 as a box map. As can be seen from the figure, these four samples are located in the areas with high iron concentrations relative to Fig. 2, which indicates that the areas with high lead concentrations might have been affected by human activities.

Spatial cluster-outlier and hotspot identification: Relative to the statistical outliers, the samples with unusual values relative to their neighbourhood are also considered to be outliers (spatial outlier, Lark 2002). As mentioned above, a series of methods have been applied and their principles are as follows: the variograms are used to model the spatial autocorrelation and with a cross-validation procedure of ordinary kriging, and an estimated value is generated for every measurement, and then the standardized estimation error can be used for identifying the spatial outlier (Laslett & Mcbratney 1990).

Among these methods, the Moran's I is a commonly used indicator of spatial autocorrelation. There are two types of Moran's I which have been reported previously: One is the global Moran's I, which is used to study the overall spatial autocorrelation, another one is LISA (local indicators of spatial association), which is applied to identify the degree of spatial autocorrelation in each specific location (Anselin 1995). More importantly, the LISA can also be used for identifying the existence of local spatial clusters by generating cluster maps (Harries 2006), which can be used for identifying the spatial hotspot and outliers (Zhang & McGrath 2004, Li et al. 2013).

Based on the calculation of GeoDa, all the samples in this study have been classified into four categories: not significant (57 samples) and significant (5 samples). Moreover, all the significant samples can be classified into three secondary categories: two, one and two samples are classified to be high-high (sample 36 and 56), low-high (sample 54) and high-low (sample 11 and 26) clusters, respectively.

According to previous studies (Zhang et al. 2008), either high-high or low-low samples can be clustered to be 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 central right of the map (sample 36) and another is located in the central west of the map (sample 56), which might be an indication of special human activities and, in the area near to these two points, the groundwater safety related to iron pollution should be careful. Moreover, the high-high, high-low and low-high samples are considered to be spatial outliers, including 5 samples as mentioned above (sample 11, 26, 36, 54 and 56).

Table 1: Summary statistics of the whole dataset and those resulting from outlier removing by three kinds of methods.

Methods	Whole data	Box plot	Spatial outlier	Combination
N of cases	62	58	57	54
N of outliers	0	4	5	8
Minimum	1.643	1.643	1.643	1.643
Maximum	272.690	136.345	272.690	136.345
Median	48.049	43.943	43.532	40.246
Mean	58.952	50.060	54.800	46.954
Standard deviation	47.219	31.133	45.563	28.883
Coefficient of variation	0.801	0.622	0.831	0.615
Skewness	2.048	0.837	2.461	0.895
p-value	<0.01	0.014	<0.01	0.043

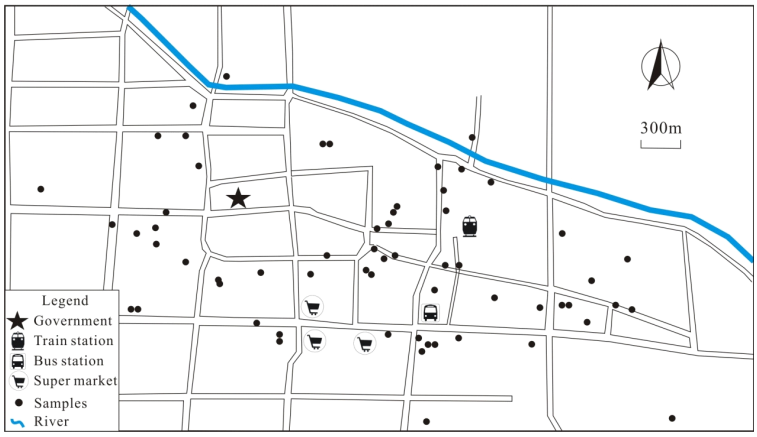

Fig. 1: Sample locations in the study area.

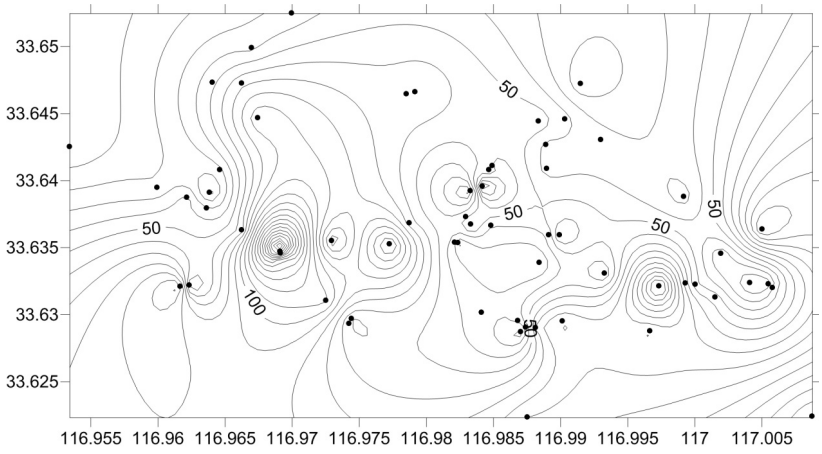

Fig. 2: Contour map of iron concentrations.

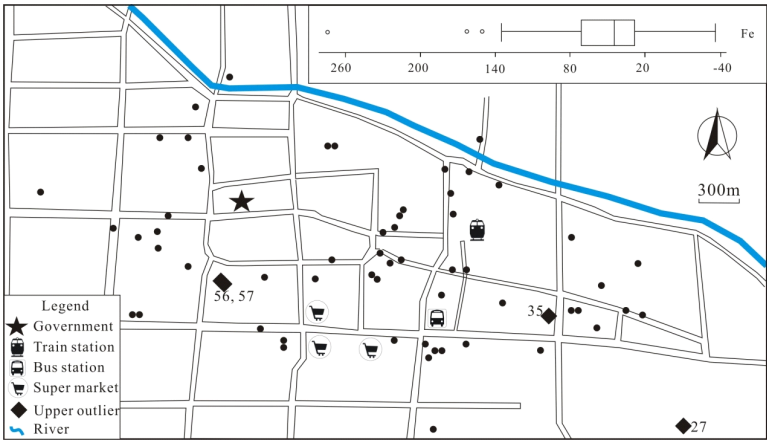

Fig. 3: Outlier distribution based on Box map.

Fig. 4: Outlier distribution based on spatial analysis

Environmental baseline evaluation: Comparatively, after statistical identification (box plot), the samples with extreme high values were removed. However, this procedure does not consider the spatial variability simultaneously, and it is therefore, inadequate to define the unique environmental baseline in a local scale. Therefore, the combination use of two methods (either box plot or spatial) can get more reliable information, as it can remove the statistical and spatial outlier simultaneously (Meklit et al. 2009).

Considering this, a total of eight samples are classified to be outliers, and the rest fifty four samples are therefore set to be environmental baseline samples and the summary statistics of the samples with different outlier identification methods are shown in Table 1.

As can be seen from the table, after outlier removing by box plot, the number of remaining samples is fifty eight, and their mean concentration is 50.060 μg/L and the p-value of normality test is 0.014, indicating that they cannot pass the normality test. However, after spatial outlier removing, the remaining samples are fifty seven with mean concentration equals to 54.800 μg/L, and they cannot pass the normality test also because they obtain p-value (<0.01) lower than the samples after box plot selection.

As to the combination method, the number of remaining samples is fifty four, and their mean concentration is 46.954 μg/L (standard deviation of 28.883 μg/L), and the environmental baseline is, therefore, established to be 0-104.720 μg/L according to Nakic et al. (2007). This result 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), the baseline values determined by the two methods are 0-81.7 and 0-96.6 μg/L, respectively. More importantly, after the combination use of box plot and spatial outlier selection also, the iron concentrations of the rest of the samples still cannot pass the normality test, as a much higher p-value (0.043) relative to other methods have been obtained (Table 1), which probably indicates that the normal distribution is suitable for the iron concentrations in the shallow groundwater of this study.

CONCLUSIONS

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

1. The groundwater samples have low iron concentrations relative to the Chinese and WHO standards. However, they have moderate-high coefficient of variation, which might be an indication of anthropogenic contribution.
2. Four outlier samples with highest iron concentrations have been identified by box plot and map, whereas five samples have been identified as outliers by spatial analysis. Among these outliers, two of them have been identified as hotpots located in the right and central west of the map, which might be an indication of special human activities.
3. The environmental baseline based on the rest of the samples is estimated to be 0-104.720 μg/L, and it is similar to the results obtained by model based objective methods.

ACKNOWLEDGEMENT

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

REFERENCES

Anselin, L. 1995. Local indicators of spatial association-LISA. Geographical Analysis, 27: 93-115.

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

Frigge, M., Hoaglin, D.C. and Iglewicz, B. 1989. Some implementations of the boxplot. The American Statistician, 43(1): 50-54.

Galuszka, A. 2007. A review of geochemical background concepts and an example using data from Poland. Environmental Geology, 52(5): 861-870.

Getis, A. and Ord, J.K. 1992. The analysis of spatial association by use of distance statistics. Geographical Analysis, 24: 189-206.

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

Ishioka, F., Kurihara, K., Suito, H., Horikawa, Y. and Ono, Y. 2007. Detection of hotspots for three-dimensional spatial data and its application to environmental pollution data. Journal of Environmental Science and Sustainable Society, 1: 15-24.

Lark, R.M. 2002. Modeling complex soil properties as contaminated regionalized variables. Geoderma, 106: 173-190.

Laslett, G.M. and Mabratney, A.B. 1990. Further comparison of spatial methods for predicting soil-pH. Soil Science Society of American Journal, 54: 1553-1558.

Li, W., Xu, B., Song, Q., Liu, X., Xu, J. and Brookes, P.C. 2013. The identification of 'hotspots' of heavy metal pollution in soil-rice systems at a regional scale in eastern China. Science of the Total Environment, 472: 407-420.

Meklit, T., Meirvenne, M.V., Verstraete, S., Bonroy, J. and 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.

Molinari, A., Guadagnini, L., Marcaccio, M. and Guadagnini, A. 2012. Natural background levels and threshold values of chemical species in three large-scale groundwater bodies in Northern Italy. Science of the Total Environment, 425: 9-19.

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

Reimann, C. and Garrett, R.G. 2005. Geochemical background-concept and reality. Science of the Total Environment, 350(1): 12-27.

Salminen, R. and Tarvainen, T. 1997. The problem of defining geochemical baselines: A case study of selected elements and geological materials in Finland. Journal of Geochemical Exploration, 60(1): 91-98.

Sun, L.H., Gui, H.R., Xu, D.S. and Huang, S.L. 2012. Heavy metal pollution in rural area of China: A case study of pond sediments from Sixian County, northern Anhui Province. Fresenius Environmental Bulletin, 21(2): 263-268.

Sun, L., Gui, H., Peng, W. and Lin, M. 2013. Heavy metals in deep seated groundwater in northern Anhui Province, China: quality and background. Nature Environment and Pollution Technology, 12(3): 533-536.

Tango, T. 1995. A class of tests for detecting 'general' and 'focused' clustering of rare diseases. Statistics in Medicine, 14: 2323-2334.

Urresti-Estala, B., Carrasco-Cantos, F., Vadillo-Pérez, I. and 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). Journal of Environmental Management, 117: 121-130.

WHO (World Health Organization) 2008. Guidelines for Drinking-Water Quality (3rd edition), Geneva: World Health Organization.

Wongsasuluk, P., Chotpantarat, S., Siriwrong, W. and Robson, M. 2014. Heavy metal contamination and human health risk assessment in drinking water from shallow groundwater wells in an agricultural area in Ubon Ratchathani province, Thailand. Environmental Geochemistry and Health, 36: 169-182.

Zhang, C., Luo, L., Xu, W. and Ledwith, V. 2008. Use of local Moran's I and GIS to identify pollution hotspots of Pb in urban soils of Galway, Ireland. Science of the Total Environment, 398: 212-221.

Zhang, C. and 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.