Source apportionment of groundwater pollution in a city’s eastern part using multivariate statistical techniques

Peng Jiang1,2, Zhenmin Ma1,2*, Ming Wen1,2

1School of resource and environment, Jinan University, Jinan 250022, China
2Research center of groundwater numerical simulation and pollution control engineering in Shandong province, Jinan 250022, China
*corresponding author’s e-mail: stu_mazm@ujn.edu.cn

Abstract: This study was carried out to assess the overall water quality and identify major chlorinated hydrocarbon variables affecting the groundwater quality. The source apportionment of groundwater pollution is important for the efficient management of groundwater resources. Based on 13 variables surveyed at 43 monitoring sites, the comprehensive application of different multivariate methods were used for determining source apportionment of groundwater chlorinated hydrocarbon pollutants in study area. Factor analysis and cluster analysis were applied to the identification of pollution sources and four potential pollution sources that explained 92.810% of the total variance were identified. The absolute principal component score-multiple linear regression was adopted to calculate the contribution of each pollution source. Regression results revealed that most variables were primarily influenced by chemical industry, electrical manufacturing, chemical fiber and agricultural source. The contributions of each pollution source to the entire study area were 43%, 32%, 14% and 11% respectively.

1. Introduction

As an important natural resource which supports our life system, groundwater is extensively utilized for drinking, irrigation, and industrial purposes. The quality of groundwater is declining due to increasing population, urbanization, industrialization, and agriculture activities[1-4]. Due to the diversity of sources of groundwater pollution, to acquire the knowledge of sources is important for effective pollution abatement [5]. Source apportionment of groundwater pollutants can enhance the comprehensive understanding to the environmental conditions and help researchers establish priorities in the sustainable groundwater management[6,7].

Based on the ambient data registered at monitoring sites, the purpose of source apportion is to identify qualitatively the type of pollutant sources, moreover, to calculate the contribution of various sources, which consequently provide a certain basis for groundwater pollution management.[8,9]. In recent years, multivariate statistics (such as cluster analysis (CA) and factor analysis (FA)) have been effectively used globally multivariate statistical method study, and used to understand clustering of sampling stations, discharge sources and their origin[10]. The study illustrates the availability of multivariate statistical techniques in the analysis and interpretation of complex data sets, the water quality assessment, and the identification of possible factors sources that influence water systems offering a valuable tool for reliable management of water resources, both in quantity and quality level [11-13]. This study indicates the necessity and usefulness of multivariate statistical techniques for the evaluation and interpretation of the data. It facilitates better information about the water quality and designs some remedial techniques to prevent future contamination[14].
Given all the above considerations, the main objectives of this work are: (1) analyzing the pollutants data and understanding the groundwater environments clearly. (2) exploring the main sources of groundwater pollutants with the assistant of factor analysis (FA) and cluster analysis (CA). (3) quantitatively calculating the contributions of each source type to each pollutant and the entire study area respectively by the absolute principal component score-multiple linear regression (APCS-MLR).

2. Materials And Methods

2.1. Sample collection and analysis
The monitoring data for this study was compiled from the groundwater pollution investigation of the typical study area. In view of the characteristics of shallow groundwater pollution system in the study area, 43 civilian wells and drilling wells were chosen as sample sites. The sample locations are shown in Fig. 1. Due to the concentration of chlorinated hydrocarbon in groundwater is low, mostly in ppb level, the test requirements of concentration of chlorinated hydrocarbon in groundwater should be controlled strictly, like test method, test condition, and test equipment. Based on China Standard Method and the U.S. Environment Protection Agency (EPA) method as the standard the concentration of chlorinated hydrocarbon in groundwater are analyzed by GC/MS (gas chromatography/mass spectrum) method. The data description was presented in table 1.

![Fig 1 The location map of study area and spatial pattern of the 43 sampling sites.](image)

**Table 1** The statistical description of measured variables in study area and “**” represents national quality standards for drinking water, China(GB 5749-2006), other’s is the reference of the United States EPA drinking water quality standards.

| Parameters | φ/% | ρ/(μg/L) |
|------------|-----|----------|
|            |     |          |
2.2. Statistical analysis
Since the elemental concentrations varied greatly among the major and trace elements, the raw data were standardized (with a mean of 0 and a standard variation of 1) before the statistical analysis.

Factor analysis (FA) is a common and convenient multivariate statistical method for data dimension reduction based on the principal components analysis (PCA) which could identify the unobservable and latent pollution sources[15]. It has gained widely application in the pollutants source apportionment. The principle of factor analysis is to interpret the internal correlations of complicated multiple variables by extracting some principal components that represent a cluster of interrelated variable[16]. Readers who are interesting to the mathematical and steps details of factor analysis can take it as reference[16,17].

Clustering analysis(CA) is a progress based on quantitative or qualitative characteristics of a large amount of experimental data for packet classification so as to understand the internal structure of a data set, and to describe each data set. The main principle of CA is that samples similar to each other in the same class, while, there has enormous diversity in different kinds of samples. Clustering analysis is normally used to verify the results of the conclusions of the factor analysis[18,19]. In this research, adopting the R cluster analysis method, and according to the similarity of variables to classify. The first step of CA is to calculate the correlation coefficient between each variable in the sample, then variables will be clustered based on the calculation results.

In the process of source apportionment, absolute principal component score-multiple linear regression(APCS-MLR) is one of the most commonly used methods in source apportionment to calculate the contribution rate of each pollution source[20]. It is important that characterizing groundwater quality and apportionment of pollution sources to groundwater pollution for efficient water management. In order to recognize and quantify the pollution sources receptor modeling by Multi-Linear Regression of the Absolute Principal Component Scores (APCS-MLR) has been used to

|                     | detection rate | Exceeding rate | Min  | Max  | Mean | quality standards |
|---------------------|----------------|----------------|------|------|------|-------------------|
| Vinyl chloride      | 93.1           | 60.34          | 0.47 | 8.53 | 3.87 | 2                 |
| Methylene chloride  | 94.83          | 94.83          | 5.21 | 16.85| 10.04| 5                 |
| 1, 2 - dichloroethane| 94.83          | 67.24          | 2.03 | 12.56| 6.6  | 5                 |
| 1, 1-2 vinyl chloride| 96.55          | 72.41          | 1.11 | 10.46| 5.34 | 7                 |
| cis,1,2-Dichloroethene| 50            | 0              | 1.12 | 2.42 | 1.95 | 50.00*            |
| trans,1,2-Dichloroethylene| 32.76      | 0              | 0.44 | 1.31 | 1.03 | 50.00*            |
| Chloroform          | 100            | 0              | 0.2  | 2.4  | 1.5  | 60.00*            |
| 1,1,1 - trichloroethane| 94.83        | 0              | 1    | 2.05 | 1.59 | 200               |
| Trichloroethylene   | 100            | 94.83          | 2.1  | 73.64| 48.87| 5                 |
| Tetrachloroethylene | 94.83          | 70.69          | 3.01 | 32.25| 10.07| 5                 |
| Carbon tetrachloride| 100            | 94.83          | 1.5  | 23.31| 10.28| 2.00*             |
| HCH                 | 15.52          | 0              | 0.1  | 0.52 | 0.37 | 5.00*             |
| DDT                 | 8.62           | 0              | 0.12 | 0.11 | 0.11 | 1.00*             |
evaluate the source apportionment of groundwater pollution [21]. After determining the number and variety of sources, APCS-MLR can quantify efficiently the contribution rate of each pollution source type [9]. In this study, APCS-MLR was employed to estimate the contribution of each pollution source to each pollutant and the entire study area.

3. Result and Discussion

3.1. Identification of pollution sources

The results of the KMO and Bartlett's Test are showed in the table 2. In general, when the KMO value is closer to 1 and significance level is closer to 0, then the correlation between variables is more stronger, which means the original variables are more suitable for factor analysis. In this test the KMO value was 0.826 and the Bartlett's Test reached extremely significant level (0.000) which express factor analysis in this article was adaptive.

| Table 2 | Results of the KMO and Bartlett's Test. |
|---------|----------------------------------------|
| Kaiser-Meyer-Olkin value of sampling adequacy. | 0.826 |
| Bartlett's test of sphericity | Test value ($\chi^2$) | 1146.832 |
| | Degree of freedom (df) | 78 |
| | Significance level (Sig.) | 0.000 |

The results of the FA are reported in the table 3. Based on the principle of eigenvalues is greater than one, the study extracted four common factors which accounted for 92.810% of the total variance. The first factor explaining 42.384% of the total variance, was strongly and positively related to Vinyl chloride, Methylene chloride, cis-1,2-Dichloroethene, trans-1,2-Dichloroethylene, Chloroform, Carbon tetrachloride. The second factor accounting for 27.464%, showed strong positive factor loadings on 1,1,1-trichloroethane, Trichloroethylene, Tetrachloroethylene. The third factor which explained 12.852% of the total variance had high loadings on 1,2- dichloroethane and 1,1-2 vinyl chloride. The fourth factor explaining 9.928% of the total variance, had positive factor loadings on HCH and DDT.

Comparing with different clustering methods, the Ward’s Method was found to be a suitable method to classify the variables. Based on the correlation efficiencies calculated by FA, the dendrogram of a cluster analysis of the 13 variables is illustrated in Fig. 2, and four distinct clusters are identified. The results of cluster analysis (CA) was consistent with the factor analysis (FA), indicating that the identification results were suitable and reliable.

| Table 3 | Total variance explained and component matrices (four common factors extracted ). |
|---------|----------------------------------------------------------------------------------|
| Component | Total variance explained |
| | Initial eigenvalues | Extraction sums of squared loadings |
| | Total | % of Variance | Cumulative (%) | Total | % of Variance | Cumulative (%) |
| 1 | 5.510 | 42.384 | 42.384 | 5.510 | 42.384 | 42.384 |
| 2 | 3.594 | 27.646 | 70.030 | 3.594 | 27.646 | 70.030 |
| 3 | 1.671 | 12.852 | 82.882 | 1.671 | 12.852 | 82.882 |
### Component matrixes

| Element                  | Component Matrix | Rotated Component Matrix |
|--------------------------|------------------|--------------------------|
|                          | Factor1 | Factor2 | Factor3 | Factor4 | Factor1 | Factor2 | Factor3 | Factor4 |
| Vinyl chloride           | 0.933   | 0.177   | 0.150   | 0.197   | 0.956   | -0.054  | -0.210  | -0.046  |
| Methylene chloride       | 0.821   | 0.467   | 0.196   | 0.131   | 0.924   | 0.243   | -0.133  | -0.133  |
| 1, 2 - dichloroethane    | -0.691  | 0.416   | -0.006  | 0.567   | -0.270  | 0.255   | 0.911   | -0.067  |
| 1, 1-2 vinyl chloride    | -0.630  | 0.388   | -0.024  | 0.647   | -0.199  | 0.182   | 0.943   | -0.070  |
| cis-1,2-Dichloroethene   | 0.936   | 0.197   | -0.039  | -0.034  | 0.835   | -0.034  | -0.401  | -0.240  |
| trans-1,2-Dichloroethylene | 0.790  | -0.259  | 0.197   | 0.424   | 0.811   | -0.437  | -0.075  | 0.236   |
| Chloroform               | 0.610   | 0.565   | 0.225   | -0.303  | 0.787   | 0.383   | -0.342  | -0.182  |
| 1,1,1 - trichloroethane  | -0.294  | 0.831   | 0.377   | -0.036  | 0.043   | 0.896   | 0.338   | -0.025  |
| Trichloroethylene        | -0.569  | 0.719   | 0.283   | -0.065  | -0.255  | 0.833   | 0.408   | -0.025  |
| Tetrachloroethylene      | -0.169  | 0.768   | 0.260   | -0.328  | -0.003  | 0.880   | 0.016   | -0.139  |
| Carbon tetrachloride     | 0.877   | 0.266   | 0.033   | 0.326   | 0.954   | -0.071  | -0.051  | -0.172  |
| HCH                      | -0.195  | -0.581  | 0.775   | 0.049   | -0.133  | -0.105  | -0.031  | 0.975   |
| DDT                      | -0.163  | -0.580  | 0.791   | 0.020   | -0.112  | -0.093  | -0.070  | 0.981   |

Extraction method: Principal Component Analysis; Rotation method: Varimax with Kaiser Normalization; Rotation converged in 5 iterations.
The variables that have high factor loading are vinyl chloride, methylene chloride, cis-1,2-Dichloroethene, trans-1,2-Dichloroethylene, chloroform and carbon tetrachloride in the first common factor (F1). These chlorinated hydrocarbons are important chemical raw materials, and they are widely used in chemical production. The vinyl chloride is an important monomer used in polymer chemistry. There is a large plastics factory in the northeast of the researched area where the groundwater has been seriously polluted. The chloroform is an important raw material in organic synthesis process. It is often used as solvents in the manufacturing process of chemical products, including freon, resin, rubber, paint, phosphorus and iodine. At the same time, it is a critical raw material for the manufacturing of dry cleaners, pesticides, floor wax, etc. As an essential chemical raw materials, the chlorinated carbon is usually used as organic solvents, fire extinguishing agent, organic chloride agent, leaching agent of spice, degreasing agent, drug extraction agent, lubricant, the fabric dry cleaners and so on. From what has been discussed above, the first common factor containing chlorinated hydrocarbon pollutants originates from chemical industry enterprises. Therefore the first common factor can be defined as chemical pollution.

The variables that have high factor loading are 1,1,1-trichloroethane, Trichloroethylene, Tetrachloroethylene in the second common factor (F2). 1,1,1-trichloroethane is a kind of incombustible solvent. It is often used as cleaning agent and metal degreasing cleaning agent. As a kind of good solvent, Trichloroethylene is commonly used as metal cleaners, metal degreasing agent and metal parts processing and surface treatment agent. Tetrachloroethylene is recyclable and it is widely used as a dry cleaning solvent for all countries in the world. From what has been discussed above, 1,1,1-trichloroethane, trichloroethylene and tetrachloroethylene are widely used in metal cleaning agent and metal degreasing agent. The second common factor can be defined as the mechanical and electrical manufacturing sources.

The variables that have high factor loading are 1, 2 - dichloroethane, 1, 1-2 vinyl chloride in the third common factor (F3). The 1, 2 - dichloroethane is more widely employed as organic solvents in industrial application. It is not only used as the raw materials of organic synthesis, but also can be used as binder, wetting agent, soaking agent, manufacturing acetyl fiberd in chemical fiber industry. The 1, 1-2 vinyl chloride is not used as solvent in the industrial production because its volatile is bigger. The 1, 1-2 vinyl chloride is used to manufacture all kinds of copolymer and compounding fiber. By comparing the profiles of 1,2-dichloroethane, with 1,1-2 vinyl chloride, the researcher found that these two kinds of chlorinated hydrocarbon pollutant concentration are more higher. So the
pollution of the 1, 2 - dichloroethane and 1, 1-2 vinyl chloride mainly come from the chemical fiber enterprises. The third common factor can be defined as chemical fiber contamination.

The variables that have high factor loading are HCH and DDT in the fourth common factor (F4). The HCH and DDT are the most typical of organochlorine pesticides that have been used in domestic production. The HCH and DDT belong to persistent organic pollutants. Properties are stable, and hard to break down even in the environment that is contaminated. What’s more, they will accumulate in the environment, which can worsen the environmental pollution. Even though the HCH and DDT have not been used in the study area, the residues of HCH and DDT in groundwater still have a certain impact on groundwater in the study area. From what has been discussed above, it is clear that the source of chlorinated hydrocarbon pollutant is pesticide pollution in the fourth common factor, so the fourth common factor can be defined as the pesticide pollution.

### 3.2. Calculation of the contributions of pollution sources

Common factor scores obtained from the factor analysis as independent variable can be took as independent variable, each variable’s standardized concentration can be seen respectively as a dependent variable, the stepwise regression method was adopted to calculate the contributions of pollution sources to each pollutant with the assistance of SPSS. The results of the APCS-MLR are presented in the table 4. Based on correlation coefficients (R2) of each APCS-MLR, the conclusion of all the APCS-MLRS were relatively accurate could be reached. The quantitative information regarding the contributions of each source type to each pollutant was calculated accurately. Most variables in the study area were primarily influenced by chemical industry (73.86%, 64.48%, 56.57%, 54.65%, 34.65% and 79.70%) of vinyl chloride, methylene chloride, cis-1,2-Dichloroethene, trans-1,2-Dichloroethylene, Chloroform and Carbon tetrachloride), electrical manufacturing(72.61%, 55.68% and 86.36% of 1,1,1 trichloroethane, trichloroethylene, tetrachloroethylene), chemical fiber (60.61% of 1, 2 - dichloroethane and 67.65% of 1, 1-2 vinyl chloride) and agricultural pollution (80.38% of HCH and 78.11% of DDT). In addition, according to the results, It could be find other unidentified sources excepting for the four pollution source types. The types of the unknown sources might need more field investigation or other feasible methods to identify.

Taking common factor scores obtained from the factor analysis as independent variables and the aggregate variables’ standardization concentrations of each sampling point as dependent variables, the same APCS-MLR method above was applied to estimate the contributions of pollution sources to the entire study area. The results are illustrated in table 5. Under the confidence level of 95%, Sig values were equal to 0.000 represented the equation was significant. R2 value was equal to 0.98, indicating that the regression model had a good fitting effect for the primary data. Consequently, the APCS-MLR was applicable and reliable. From the table the regression equation was calculated as following: Z = 4.41F1 + 3.08F2 + 1.30F3 + 1.09F4 . This formula showed that contributions of chemical industry, electrical manufacturing, chemical fiber and agricultural source to pollutions of the entire study area were 43%, 32%, 14% and 11% respectively. The contribution of the chemical industry was the highest, the electrical manufacturing took the second place, the chemical fiber and agricultural sources were following subsequently.

**Table 4 Source contribution to each variable.**

| Parameters            | Potential pollution source | USa | R²  |
|-----------------------|----------------------------|-----|-----|
|                       | F1            | F2      | F3      | F4      | USa |     |
| Vinyl chloride        | 78.36         | 4.43    | 17.21   |         |     | 0.96|
| Methylene chloride    | 64.48         | 16.96   | 9.28    | 9.28    |     | 0.95|
| 1, 2 - dichloroethane | 17.96         | 16.97   | 60.61   | 4.46    |     | 0.97|
Table 5 Regression coefficients.

| Model | Unstandardized Coefficients | Standardized Coefficients | t    | Sig. | R²   |
|-------|-----------------------------|----------------------------|------|------|------|
|       | B                           | Std. Error                 | Beta |      |      |
| Constant | 9.310E-6                   | 0.03                       | 0.00 | 1.00 |      |
| F1     | 4.14                        | 0.03                       | 0.76 | 124.12 | 0.00 |
| F2     | 3.08                        | 0.03                       | 0.57 | 92.31 | 0.00 |
| F3     | 1.30                        | 0.03                       | 0.24 | 39.04 | 0.00 |
| F4     | 1.09                        | 0.03                       | 0.20 | 32.79 | 0.00 |

Sig.: Significant level. R²: coefficient of determination.

4. Conclusions
In this study, multivariate statistical methods like factor analysis (FA), cluster analysis (CA) and absolute principal component score-multiple linear regression (APCS-MLR) were adopted to identify the different groundwater pollution source types and calculate the contribution of each type of pollution source in study area, using 43 sampling sites data. Results showed that the comprehensive application of FA, CA and APCS-MLR was effective for groundwater pollution source apportionment.

Through the analysis of data of 43 sampling sites, we could draw a conclusion that the groundwater in the study area was polluted by chlorinated hydrocarbon. According to the results of the FA and CA, the pollution sources could be grouped into four categories, namely chemical industry, electrical manufacturing, chemical fiber, and agricultural source, explaining 92.810% of the total variances.

The calculation of each pollution source based on APCS-MLR revealed the contribution rates of these four pollution sources to the entire study area were respectively 43%, 32%, 14% and 11%. However, it exists another latent sources in the study area. Therefore, more field investigations or other feasible methods need be required to have a further identification of the latent pollution sources.
Acknowledgments
This work was financially supported by Shandong provincial key research project (no.2015GSF117025) and Shandong provincial natural science foundation, China (no. ZR2014DM011).

References
[1] Hani, H (1990). The analysis of inorganic and organic pollutants in soil with special regard to their bioavailability. Int J Environ Anal Chem, 39(2):197–208.
[2] Kumar, N., (2010). Evaluation of Groundwater quality in shallow and deep aquifers: a case study. Report and Opinion. 2(9):75–87.
[3] Pangarkar, B.L., Sane, M.G and Mahendra, G. (2011). Reverse osmosis and membrane distillation for desalination of groundwater: a review. ISRN Materials Science, 2011(2090-6080).
[4] Chidambaram, S., Prasad, M.B.K and Prasanna, M.V. (2014). Evaluation of Metal Pollution in Groundwater in the Industrialized Environments in and Around Dindigul, Tamilnadu, India. Water Quality Exposure & Health, 7(3):1-11.
[5] Xue, L., Lang, Y. and Liu, A., et al (2010). Application of CMB model for source apportionment of polycyclic aromatic hydrocarbons (PAHs) in coastal surface sediments from Rizhao offshore area, China. Environmental monitoring and assessment, 163(1-4), 57-65.
[6] Kolovos, A., Christakos, G., and Serre, M.L., et al (2002). Computational BME solution of a stochastic advection reaction equation in the light of site-specific information. Water Resources Research, 38 (12), 1318–1334.
[7] Huang, F., Wang, X. and Lou, L., et al (2010). Spatial variation and source apportionment of water pollution in Qiantang River (China) using statistical techniques. Water Research, 44(5), 1562-1572.
[8] Tauler, R., Viana, M., and Querol, X., et al (2009). Comparison of the results obtained by four receptor modelling methods in aerosol source apportionment studies. Atmospheric Environment, 43(26), 3989-3997.
[9] Cai, W.L., Luo, G., and Xiao-Yi, X.U., et al (2012). Contamination Characteristics of Polycyclic Aromatic Hydrocarbons (PAHs) in Surface Water from Jialing River in Chongqing. Environmental Science, 33(7):2341-6.
[10] Varol, M., Gökot, B. and Bekleyen, A., et al (2012). WATER QUALITY ASSESSMENT AND APPORTIONMENT OF POLLUTION SOURCES OF TIGRIS RIVER (TURKEY) USING MULTIVARIATE STATISTICAL TECHNIQUES—A CASE STUDY. River Research & Applications, 28(9):1428–1438.
[11] Muangthong, S. and Shrestha, S. (2015). Assessment of surface water quality using multivariate statistical techniques: case study of the Nampong River and Songkhram River, Thailand. Environmental Monitoring & Assessment, 187(9):1-12.
[12] Khound, N.J. and Bhattacharyya, K.G., (2016). Multivariate statistical evaluation of heavy metals in the surface water sources of Jia Bharali river basin, North Brahmaputra plain, India. Applied Water Science, 1-10.
[13] Hema, S., Subramani, T. and Elango, L., (2014). Assessment of Surface Water Quality Using Multivariate Statistical Techniques in a Part of River Cauvery, Tamil Nadu, India. Journal of Environmental Science & Engineering, 56(3):277-282.
[14] Su, S., Zhi, J., and Lou, L., Huang, F., et al (2011). Spatio-temporal patterns and source apportionment of pollution in Qiantang River (China) using neural-based modeling and multivariate statistical techniques. Physics and Chemistry of the Earth, Parts A/B/C, 36(9-11), 379-386.
[15] Li, X., Lee, S. L. and Wong, S. C., et al (2004). The study of metal contamination in urban soils of Hong Kong using a GIS-based approach. Environmental Pollution, 129(1), 113-124.
[16] Pekey, H., Karakas, D. and Bakoglu, M., (2004). Source apportionment of trace metals in surface waters of a polluted stream using multivariate statistical analyses. Marine Pollution Bulletin, 49(9), 844-850.
[17] Yidana, S. M., Banoeng-Yakubo, B. and Akabzaa, T. M. (2010). Analysis of groundwater quality using multivariate and spatial analyses in the Keta basin, Ghana. *Journal of African Earth Sciences*, 58(2), 220-234.

[18] Wang, X. S. and Qin, Y. (2006). Environmental rise and sources of heavy metals in Xuzhou urban topsoil. *Geochimica*, 35(1), 88-94.

[19] Facchinelli, A., Sacchi, E. and Mallen, L. (2001). Multivariate statistical and GIS-based approach to identify heavy metal sources in soils. *Environmental Pollution*, 114(3), 313-324.

[20] Zhou, F., Huang, G. H. and Guo, H., et al. (2007). Spatio-temporal patterns and source apportionment of coastal water pollution in eastern Hong Kong. *Water Research*, 41(15), 3429-3439.

[21] Gulgundi, M. S. and Shetty, A. (2016). Identification and Apportionment of Pollution Sources to Groundwater Quality. *Environmental Processes*, 3(2):1-11.