Multivariate statistical approaches to identify the major factors governing groundwater quality

Tao Chen¹ · Huafei Zhang¹ · Chengxun Sun¹ · Hongyan Li¹ · Yang Gao¹

Received: 24 December 2016 / Accepted: 1 October 2018 / Published online: 31 October 2018
© The Author(s) 2018

Abstract

Multivariate statistical techniques, discriminant analysis, cluster and principal component analysis were applied to the dataset on groundwater quality of Longyan basin of Fujian Province (South China), to extract principal factors controlling the source variations in the hydrochemistry and identify the major factors affecting groundwater quality. The dataset covers ten parameters of monitored wells at five typical locations in the region. The results were evaluated in accordance with the groundwater quality standards suggested by Specification GB/T14848-93, “The Quality Standard of Underground Water.” Cluster analysis results reveal that the groundwater in the study area is classified into two groups (A: 2000–2007 and B: 2008–2011) between the sampling sites, reflecting regular characters of interannual variability. Factor analysis/principal component analysis, applied to the datasets of the two different groups obtained from cluster analysis, resulted in three factors accounting for 85.5% and 100% of the total variance in the water quality datasets, respectively. Three of the ten parameters processed by discriminant analysis obtained a conformation rate of 100% which allowed a reduction in the dimensionality of large dataset, and also it found that most discriminant parameters (total alkalinity, chloride ion, sulfate ion) are responsible for temporal variation of water quality. So this study illustrates the usefulness of multivariate statistical techniques for interpreting complex datasets of water quality, identifying pollution sources/factors for effective groundwater quality management.

Keywords Multivariate statistics · Factor analysis · Cluster analysis · Groundwater quality

Introduction

Sustainable socioeconomic development of every community depends greatly on the sustainability of the available water resources. Rapidly shrinking surface water resources due to over-exploitation and subsequently resulted contamination with several chemical and biological agents all over the globe have shifted tremendous pressure on the groundwater resources. Groundwater serves as the primary water supply in most arid or semiarid area, which is particularly valuable for the human living due to its relative cleanness. However, it is difficult and costly to remediate once it is polluted. The groundwater quality is affected by many factors, such as the lithology of the water-bearing sediments, climatic conditions, inputs from the adjacent water bodies

and anthropogenic activities. For the decades, researches have been focusing on the hydrochemical analysis in various ways. The Piper diagram has been applied broadly to investigate the groundwater facies for further research such as revealing the evolution of phreatic water and understanding the hydrochemical characteristics as well as the formation mechanism of the groundwater (Yang et al. 2016a, b, 2017). However, it is difficult to study the inherent relationship and the interaction of these values by the common methods; some more sophisticated data analysis techniques are required to interpret groundwater quality effectively. The application of different multivariate statistical techniques, such as cluster analysis (CA), principal component analysis (PCA), factor analysis (FA) and discriminant analysis (DA), has received substantial attention in the study of water quality, which can help to interpretate the complex data matrices to better understand the water quality and ecological status of the studied systems, allows the identification of possible factors/sources that influence water systems and offers a valuable tool for reliable management of water resources as well as rapid solution to pollution problems. Recent studies

✉ Tao Chen
cscct@126.com

¹ Electric Power Research Institute, Jilin Electric Power Co., Ltd of State Grid, Changchun 130021, People's Republic of China

have confirmed the usefulness of these techniques which have been employed to extract critical information from hydrochemical datasets with respect to groundwater quality assessment in several studies (Paopatheodorou et al. 2007; Belkhir et al. 2010; Rao et al. 2010; Rao 2014; Yang et al. 2015a, b).

In the present study, a large data matrix, obtained during a 12-year (2000–2011) monitoring program, was analyzed with different multivariate statistical techniques to identify water quality variables responsible for spatial and temporal variations and the controlling processes of groundwater quality and to interpret the relative importance of the chemical variables and the influence of possible sources (natural and anthropogenic) on the water quality parameters of Longyan basin, a coastal area in Fujian Province, China.

Materials and methods

Monitored parameters

The groundwater samples were collected from the monitoring wells of Longyan City Water Bureau in wet season and dry season from 2000 to 2011. We chose the most complete 5 monitoring wells of all of 60 wells in different locations of the study area. Ten parameters, pH, total alkalinity (T-Alk), total hardness (T-Hard), total dissolved solids (TDS), calcium ion (Ca^{2+}), sodium ion (Na^+), potassium ion (K^+), chloride ion (Cl^-), sulfate ion (SO_4^{2-}) and nitrite nitrogen ($\text{NO}_3\text{-N}$), were analyzed. The unit of these parameters is mg/L except pH.

Analytical methods

In this study, three multivariate techniques, hierarchical cluster analysis (CA), discriminant analysis (DA) and principal component analysis/factor analysis (PCA/FA), were employed to analyze the temporal variations of the selected parameters. All mathematical and statistical computations were performed using Microsoft Excel and SPSS 19.0.

Hierarchical cluster analysis (HCA)

Cluster analysis is a statistical tool to classify the true groups of data according to their similarities to each other. A number of studies using these techniques to successfully classify water samples have been conducted (Arslan 2013; Dash 2006). Hierarchical cluster analysis is the major method for finding relatively homogeneous cluster of cases based on measured characteristics. It starts with each case as a separate cluster, i.e., there are as many clusters as cases, and then combines the clusters sequentially by reducing the number of clusters at each step until only one cluster is left (Yang et al. 2015a,

b). In this study, first of all, using “average value is 1” as transform values can standardize parameters to eliminate the effect of index dimension, using squared Euclidean distance as a proximity measure, and group average algorithm was used to classify the data into different groups.

Discriminant analysis (DA)

Discriminant analysis is the most widely used multivariate statistical technique, whose basic idea is summarizing the regular rule of the data and then establishing discriminant function to classify the new data into different groups. The DA allows the differences between groups (clusters) to be studied in respect of several predictor variables of water quality parameters simultaneously. Therefore, discriminant analysis is generally used for two purposes: the first purpose is the description of group separation in which the linear functions of several variables are used to describe the differences between the groups and to identify the relative contribution of all variables to the separation of groups, and the other purpose is the prediction or allocation of new observations to groups in which linear or quadratic functions of the variable are used to assign an observation to one of the groups (Majid et al. 2013; Yang et al. 2015a). In this study, the criterion function for verifying clusters by CA was established through stepwise discriminant analysis, using entry F value as 3.84 and removal F value as 2.71. Stepwise discriminant analysis proved a good method for classifying water quality clusters and had the advantage of allowing the percentage of correctly classified results to be tested using cross-validation.

Principal component analysis/factor analysis (PCA/FA)

PCA is a statistical data reduction tool which can be used to aggregate the effects of many variables into a small subset of factors, to interpret observed relationships among variables, to yield simpler relationships that provide insight into the underlying structure of the variables and to assess controls on groundwater composition (Liu et al. 2003). It assesses the associations between variables as it indicates the participation of individual chemicals among several factors of influence (Mohaparta et al. 2011; Venkatesh et al. 2013).

Both of the principal component analysis and factor analysis look for linear combination of variables which best explain the data. The principal components (PCs) are weighted linear combinations of the original variables. PC provides information on the most meaningful parameters, which describe the whole dataset while affording data reduction with a minimum loss of original information (Iscen et al. 2008). This study retains only factors with eigenvalues that exceed 1. Factors with eigenvalues > 1 explained more total variation in the data than individual groundwater

quality variables, and factors with eigenvalue < 1 explained less total variation than individual variables. It comes out varimax rotation's principal component score in virtue of SPSS statistical software, and integrated score is calculated. And the various pollution sources were determined by factor analysis method.

Results and discussion

Groundwater quality characteristics

A statistical summary of hydrochemical parameters such as minimum, maximum, average, standard deviation, variable coefficient is given in Table 1. In ten measured parameters of the 12 years in comparison with the national guidelines of China for the entire study area, all the parameters were lower than the national standard II. The overall groundwater of the five samples was relatively suitable for drinking, domestic, agricultural and industrial purposes. The pH values of all the groundwater samples range between 7.46 and 7.86 with mean of 7.63, indicating that the groundwater was neutral to slightly alkaline, well within the acceptable limit of 6.5–8.5 (GB/T 14848-93). Total hardness varies from 158.56 to 202.10 mg/L, with a mean of 177.66 mg/L, which shows that water is safe for drinking purpose. According to Specification GB/T14848-93, the hardness up to 75 mg/L is classified as soft water, 76–150 mg/L as moderately soft water, 151–300 mg/L as hard, and more than 300 mg/L as very hard water. On this basis, the results showed that all the samples were hard. The ratio of T-Hard to T-Alk can be used to evaluate whether or not rock dissolution is influenced by artificial acids. Chemical analysis of water samples indicated that the order of these ionic concentrations was $\text{T-Alk} > \text{Ca}^{2+} > \text{SO}_4^{2-} > \text{Cl}^- > \text{Na}^+ > \text{K}^+$, which explains that Ca^{2+} was the dominant cation and SO_4^{2-} was the dominant anion, the main groundwater hydrochemical type of Longyan

Table 1 Summary statistics of chemical constituents of groundwater during 12 years (2000–2011)

Parameters	Min.	Max.	Avg.	SD	Cv. (%)
pH	7.46	7.86	7.63	0.12	1.59
T-Alk	151.98	173.85	159.35	5.75	3.61
TH	158.56	202.10	177.66	12.69	7.14
TDS	179.86	239.11	212.74	18.26	8.58
Ca^{2+}	56.91	72.05	63.32	4.47	7.06
K^+	0.92	1.62	1.19	0.21	17.49
Na^+	4.33	7.66	5.28	1.03	19.43
Cl^-	4.01	10.02	6.33	1.65	26.07
SO_4^{2-}	12.00	23.68	16.94	4.26	25.13
NO_3^- -N	0.71	1.29	0.90	0.16	17.44

basin is $\text{HCO}_3\text{-Ca}$, several samples are $\text{HCO}_3\text{-SO}_4\text{-Ca}$ or $\text{HCO}_3\text{-Na-Ca}$ type, and this study area has low total dissolved solids. High values of variable coefficient suggested that the uneven distribution of the long-term spatiotemporal variations of these parameters is due to the high discrete degree, K^+ (17.49%), Na^+ (19.43%), Cl^- (26.07%), SO_4^{2-} (25.13%), $\text{NO}_3\text{-N}$ (17.44%); sometimes we can eliminate these singular points in which parameter's variable coefficient is over 15%.

Cluster analysis in water quality at the interannual scale

In this study, CA classified the temporal water quality data into different clusters. The dendrogram is presented in Fig. 1, grouping 12 years into two clusters at $D_{\text{link}}/D_{\text{max}} < 15$; Group A (GA) included 2000–2008, and Group B (GB) included 2009–2011. The results indicate that CA was able to classify the temporal water quality in an optimal way and offer a reliable classification of groundwater quality in the whole study area. Furthermore, CA was useful for demonstrating regular temporal differences between the two groups.

Boxplot of Groundwater Quality Index

The temporal variation of water quality mainly depends on the variation of the different water quality parameters. To study the temporal variation of water quality between the two groups, each water quality parameter of the two groups is shown in Fig. 2. The boxplots of the ten constituents showed a relatively obvious increase in concentration median between Group A and Group B. A clear increase with median pH is found in Group B compared to Group A. In conclusion, we note that the groundwater quality

Fig. 1 Dendrogram of interannual clustering results ($D_{\text{link}}/D_{\text{max}} < 15$)

Fig. 2 Boxplot of Water Quality Index for the two clustered groups

was partial to slightly alkaline and the pH values had interannual variations. Ca^{2+} is the major cation of the low TDS groundwater, mainly derived from dissolved sediment of carbonate rock and weathering dissolution of calcic minerals in magmatite and metamorphic rocks (Isen et al. 2008). Na^+ was higher in Group B compared to the median recorded in Group A. The concentration of Na^+ of GA was more stable than GB according to the lower and upper boundaries of whisker; the increase in Cl^- depends usually on the increase in TDS; and the concentration of Cl^- can be used to describe the evolution of groundwater. As the distance of flow path increases, the concentration will also increase. In addition, the SO_4^{2-} has been significantly increased, and no obvious temporal variation of GB can be explained by the bound of whisker which is short. The trend of NO_3^- was indistinctively rising, maybe due to agricultural pollution or industrial pollution.

Discriminant analysis in water quality at the interannual scale

The discriminant analysis was performed using water quality parameters as predictors of membership in streams of water quality groups obtained by CA (Majid et al. 2013). DA at temporal scale was performed on the raw data after dividing the whole dataset into two interannual groups. The statistical summary of the DA is shown in Table 2; Wilks' lambda and Chi-square distributions were used to test the significance of discriminant function (Table 2). CA clustered the water quality of 12 years into two groups, so there're 1 unstandardized discriminant function (DF) and 2 standardized (Fisher's linear) discriminant functions. The unstandardized coefficients generated from DA were used for computing the discriminant score, whereas the standard coefficients were used to assess the contribution of each independent variable to the discriminant function. The test of unstandardized DA function refers

Table 2 Statistical summary for the discriminant function

Eigenvalue	% of variance	Cumulative (%)	Canonical correlation
822.042	100.0	100.0	0.999
Wilks' lambda	Chi-square	df	Sig.
0.031	29.391	3	0.000

to the statistical summary in Table 2. The statistical summary of Table 2 accounted for 100% of the total variances, and the statistical summary can be used to test the significant effect of all of the discriminant functions; small Wilks' lambda and high Chi-square value indicate high significance of the discriminant function, whereas high eigenvalues corresponding to high canonical correlation indicate effectiveness of the discriminant function in differentiating between the cases. Small sigma indicated significance level; it concludes that these DFs were effective because significance level is below 0.01.

The DF coefficients generated in the stepwise discriminant analysis for three significant parameters (T-Alk, Cl^- and SO_4^{2+}) are shown in Table 3. The value of DF coefficients measures the effectiveness of these parameters; high absolute value of coefficient indicated that this parameter has held the important position in discriminant analysis. The DF generated is given in the following equations:

$$\text{DF1} = 0.99[\text{T-Alk}] - 3.085[\text{Cl}^-] + 1.184[\text{SO}_4^{2+}] - 158.217$$

$$\text{DF2} = 172.698[\text{T-Alk}] - 546.650[\text{Cl}^-] + 181.610[\text{SO}_4^{2+}] - 13,290.467$$

$$\text{DF3} = 183.324[\text{T-Alk}] - 579.775[\text{Cl}^-] + 194.326[\text{SO}_4^{2+}] - 15,008.574$$

where DF1, DF2, DF3 are the discriminant scores and T-Alk, Cl^- , SO_4^{2+} are the independent variables. This indicates

that the main contributing parameters to the equation are T-Alk, Cl^- , SO_4^{2+} , which suggested that these parameters are important for differentiation among the groups.

All of the posterior probability of the computational classification was above 100%. It's effective for discriminant procedure that posterior probability should be greater than 90%, improving the stability of the DF and the reliability of the discriminant results. In conclusion, it can indicate that the cluster result of CA was correct and reliable, proving that the groundwater quality had significant temporal variation in the study area.

The results of principal component analysis/factor analysis

PCA/FA was performed on the correlation matrix of rearranged datasets separately for two temporal clusters to infer the possible influences and compare the compositional patterns of analyzed water samples. The rotated loading matrix, eigenvalues for different factors, percentage variance explained and cumulative percentage of variance are signalized in Table 4. The PCA of the annual mean concentrations of the water quality cons identified factors with eigenvalues > 1.

In the Group A, three independent factors were extracted which explained 85.503% of the total variance. The first varifactor explained 43.8% of the total variance in the dataset, with strong positive loadings on Ca^{2+} (94.2%), T-Hard (94.1%), SO_4^{2-} (88.1%), TDS (80.1%), K^+ (76.5%), and these elements are related to each other. TDS consists of the concentration of Ca^{2+} and Mg^{2+} , and Ca^{2+} is the major ion which influences the value of TDS. VF1 shows that the groundwater quality of this area has a tendency to be harder with time. Ca^{2+} is negatively correlated with Na^+ , which is shown in VF1, indicating that Ca^{2+} and Na^+ were not

Table 3 Stepwise discriminant function coefficients

Parameters	Unstandardized DF coefficients	Standardized DF coefficients	
	DF1	DF2	DF3
T-Alk	0.99	172.70	183.32
Cl [−]	−3.09	−546.65	−579.78
SO ₄ ²⁺	1.18	181.61	194.33
Constant	−158.22	−13,290.47	−15,008.57

Table 4 PCA results summarized the rotated component matrix of standardized water quality data and the eigenvalues of each PC

Parameters	Group A			Group B		
	VF1	VF2	VF3	VF1	VF2	VF3
pH	0.345	0.826	0.225	-0.509	0.150	0.847
T-ALK	0.172	0.073	0.852	0.962	-0.170	0.213
TH	0.941	0.091	0.276	0.970	-0.112	-0.216
TDS	0.801	-0.120	0.514	0.382	0.229	0.895
Ca^{2+}	0.942	0.135	0.241	0.962	-0.198	-0.190
K^+	0.765	0.374	-0.125	0.235	0.951	0.200
Na^+	-0.156	-0.774	0.259	-0.116	0.854	0.506
Cl^-	0.654	0.269	0.675	0.984	0.176	0.006
SO_4^{2-}	0.881	0.281	0.126	-0.563	0.791	-0.239
NO_3^- -N	0.002	0.851	0.315	-0.406	0.690	0.600
Eigenvalue	4.377	2.344	1.829	4.716	2.923	2.361
% Variance explained	43.8	23.4	18.3	47.2	29.2	23.6
% Cumulative variance	43.8	67.2	85.5	47.2	76.4	100

from the same source. VF2 explained 23.4% of the total variance, being strongly represented by pH and moderately represented by nitrate nitrogen. This factor indicated that the variation of acid–base depends on nitrate nitrogen, due to artificial pollution, agricultural pollution or industrial pollution. The negative correlation with Na^+ can be attributed to the temporal change. VF3 explaining 18.3% of the total variation has a strong positive loading on T-Alk and a strong negative loading on K^+ .

For the dataset regarding Group B, three factors explained all of the total variance. The first factor explained 47.2% of the total variance, with significant positive loadings on T-Alk, T-Hard, Ca^{2+} , Cl^- and negative loadings on pH, Na^+ , SO_4^{2-} . Na^+ has negative correlation with Cl^- , indicating that the solution of rock salt is slight in groundwater. VF2 explained 29.2% of the total variance, with significant positive loadings on K^+ , Na^+ , SO_4^{2-} and negative correlation with T-Alk, T-Hard and Ca^{2+} , indicating Na^+ and K^+ derived from the same source. VF3 explained 23.6% of total variance, with significant positive loadings on pH and TDS and negative correlation with T-Hard, Ca^{2+} , SO_4^{2-} .

Conclusions

Multivariate statistical techniques, hierarchical cluster analysis, discriminant analysis and principal component/factor analysis have successfully been used to derive information from the dataset about the possible influences of the environment on groundwater quality and identify the factors or sources responsible for water quality variations and hydrochemical characterization in a coastal aquifer, South China. These methods are important to avoid misinterpretation of environmental monitoring data due to uncertainties. Cluster analysis revealed two different groups (A: 2000–2007 and B: 2008–2011) of similarities between the sampling sites, reflecting regular characters of interannual variability. Three of the ten parameters processed by discriminant analysis obtained a conformation rate of 100% which helped in data reduction, and also it found that most discriminant parameters (total alkalinity, chloride ion, sulfate ion) are responsible for temporal variation of water quality. Principal component analysis of the two different groups resulted in three factors accounting for 85.503% and 100% of the total variance in the water quality datasets, respectively.

Acknowledgements This research was financially supported by National Natural Science Foundation of China (Grant Number 41402202).

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate

credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Arslan H (2013) Application of multivariate statistical techniques in the assessment of groundwater quality in seawater intrusion area in Bafra Plain, Turkey. *Environ Monit Assess* 185(3):2439–2452
- Belkhir L, Boudoukha A, Mouni L, Baouz T (2010) Multivariate statistical characterization of groundwater quality in Ain Azel plain, Algeria. *Afr J Environ Sci Technol* 4(8):526
- Dash MK (2006) Assessment of water quality using multivariate statistical techniques in the coastal region of Visakhapatnam, India. *Environ Monit Assess* 186(10):6385–6402
- Isçen CF, Özgür E, İlhan S et al (2008) Application of multivariate statistical techniques in the assessment of surface water quality in Uluabat Lake, Turkey. *Environ Monit Assess* 144(1–3):269–276
- Liu CW, Lin KH, Kuo YM (2003) Application of factor analysis in the assessment of groundwater quality in a blackfoot disease area in Taiwan. *Sci Total Environ* 313:77
- Majid A, Abdullah RB, Mohd Kamil Y et al (2013) Multivariate statistical techniques for the assessment of seasonal variations in surface water quality of pasture ecosystems. *Environ Monit Assess* 185(10):8649–8658
- Mohaparta PK, Vijay R, Pujari PR, Sundaray SK, Mohanty BP (2011) Determination of processes affecting groundwater quality in the coastal aquifer beneath Puri city, India: a multivariate statistical approach. *Water Sci Technol* 64(4):809
- Paopatheodorou G, Lambrakis N, Panagopoulos G (2007) Application of multivariate statistical procedures to the hydrochemical study of a coastal aquifer: an example from Crete, Greece. *Hydrol Process* 21:1482
- Rao NS (2014) Spatial control of groundwater contamination, using principal component analysis. *J Earth Syst Sci* 123(4):715
- Rao YR, Keshari AK, Gosain AK (2010) Evaluation of regional groundwater quality using PCA and geostatistics in the urban coastal aquifer, East Coast of India. *Int J Environ Waste Manag* 5(1–2):163
- Venkatesh U, Vivekanand H, Hernandez EA (2013) Assessment of groundwater water quality in central and southern Gulf Coast aquifer, TX using principal component analysis. *Environ Earth Sci*. <https://doi.org/10.1007/s12665-013-2896-8>
- Yang Q, Zhang J, Wang Y, Fang Y, Martín JD (2015a) Multivariate statistical analysis of hydrochemical data for shallow groundwater quality factor identification in a coastal aquifer. *Pol J Environ Stud* 24(2):769–776
- Yang Q, Hou Z, Wang Y, Zhao Y, Delgado J (2015b) A comparative study of shallow groundwater level simulation with WA-ANN and ITS model in a coastal island of south China. *Arab J Geosci* 8:6583–6593
- Yang Q, Wang L, Ma H, Yu K, Martin JD (2016a) Hydrochemical characterization and pollution sources identification of groundwater in Salawusu aquifer system of Ordos Basin, China. *Environ Pollut* 216:340–349
- Yang Q, Li Z, Ma H, Martin JD (2016b) Identification of the hydrogeochemical processes and assessment of groundwater quality using classic integrated geochemical methods in the Southeastern Part of Ordos basin, China. *Environ Pollut* 218:879–888
- Yang Q, Zhang L, Hou Z, Lei X, Tai W, Chen W, Chen T, Martin JD (2017) Shallow groundwater quality assessment: use of the improved Nemerow pollution index, wavelet transform and neural networks. *J Hydroinform* 19(5):784–795

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