Front. Environ. Sci. Eng. 2014, 8(6): 895–904
DOI 10.1007/s11783-014-0736-z

RESEARCH ARTICLE

Assessment of temporal and spatial variations in water quality using multivariate statistical methods: a case study of the Xin'anjiang River, China

Xue LI[1], Pengjing LI[1], Dong WANG[2], Yuqiu WANG (✉)[1]

1 College of Environmental Science and Engineering, Nankai University, Tianjin 300071, China
2 Chinese Academy for Environmental Planning, Water Environment Institute, Beijing 100012, China

Abstract This study evaluated the temporal and spatial variations of water quality data sets for the Xin'anjiang River through the use of multivariate statistical techniques, including cluster analysis (CA), discriminant analysis (DA), correlation analysis, and principal component analysis (PCA). The water samples, measured by ten parameters, were collected every month for three years (2008–2010) from eight sampling stations located along the river. The hierarchical CA classified the 12 months into three periods (First, Second and Third Period) and the eight sampling sites into three groups (Groups 1, 2 and 3) based on seasonal differences and various pollution levels caused by physicochemical properties and anthropogenic activities. DA identified three significant parameters (temperature, pH and *E.coli*) to distinguish temporal groups with close to 76% correct assignment. The DA also discovered five parameters (temperature, electricity conductivity, total nitrogen, chemical oxygen demand and total phosphorus) for spatial variation analysis, with 80.56% correct assignment. The non–parametric correlation coefficient (Spearman R) explained the relationship between the water quality parameters and the basin characteristics, and the GIS made the results visual and direct. The PCA identified four PCs for Groups 1 and 2, and three PCs for Group 3. These PCs captured 68.94%, 67.48% and 70.35% of the total variance of Groups 1, 2 and 3, respectively. Although natural pollution affects the Xin'anjiang River, the main sources of pollution included agricultural activities, industrial waste, and domestic wastewater.

Keywords Xin'anjiang River, multivariable statistical analysis, temporal variation, spatial variation, water quality

Received December 16, 2013; accepted May 8, 2014

E-mail: yqwang@nankai.edu.cn

1 Introduction

Most Chinese rivers and groundwater sources have poor and declining water quality. Industrial and municipal wastewater discharges cause widespread water pollution. Agricultural run–offs, including fertilizers, pesticides, manure, and increasing amounts of wastewater, also cause water pollution [1]. According to the data published by the Chinese Academy of Science in 2007: two–thirds of the 669 cities in China have water shortages, more than 40% of China's rivers are severely polluted, 80% of China's lakes suffer from eutrophication, and about 300 million rural residents in China lack access to safe drinking water [2].

To begin to improve water quality, China needs to build an integrated network to monitor surface water and groundwater, and then use the network to assess and set water policies through an integrated water–resource management system [3]. According to the Regulation for Water Environmental Monitoring [4], the water quality of all the major river systems in China is regularly monitored at several sites for a great number of physicochemical, bacteriological and hydrological parameters, leading to the creation of tremendous and complex databases. Therefore, it is necessary to optimize the monitoring network so that it recognizes the representative parameters and extracts only the most meaningful information from large, complicated data sets, without missing any useful information. A model that combined end–member mixing with Principal Component Analysis (PCA) was developed to estimate mixing proportions through computation and knowledge of a given hydrological system [5–7]; however, the model cannot be adapted for our study given the obtained sets of data. Multivariate statistical and exploratory data analysis techniques are appropriate tools for the treatment of analytical and environmental data and are currently being

used more frequently in experiments [7–12].

The objective of this study is to determine seasonal and spatial variations in the quality of the surface water of the Xin'anjiang River. This study also aims to examine the impact of physical and climatic basin characteristics on typical water quality parameters. In recent years, multivariate statistical techniques, including cluster analysis (CA), principal component analysis (PCA) and discriminant analysis (DA), have been used widely to assess surface water quality. Multivariate statistical techniques have also been used to evaluate spatial or temporal variations caused by natural or anthropogenic factors and possibly linked to seasonality [13]. Multivariate statistical techniques were used to examine the large data set obtained from the Xin'anjiang River to determine the spatio–temporal distributions of water contamination and to ascertain the correlations between basin characteristics and water quality parameters in a GIS environment. The results could help to deduce the potential pollution sources for the Xin'anjiang River and could provide a valuable method for water quality agencies to effectively focus their resources on severe water pollution at the watershed scale.

2 Data and methods

2.1 The study area

The Thousand Island Lake was formed by the construction of the Xin'anjiang dam in 1958. When water in the dam reached its highest level (108 m), an area of approximately 580 km^2 was inundated and 1,078 islands larger than 0.25 ha were created out of former hilltops, creating the Thousand Island Lake [14]. The Thousand Island Lake is a large artificial reservoir, located in the western Zhejiang Province (29°22′–29°50′ N, 118°34′–119°15′ E), created for the purpose of generating hydroelectricity [15]. The economic development and drinking water safety of the Zhejiang Province, especially Hangzhou City, is closely connected to the water quality of this lake. This study focuses on the Xin'anjiang River, not only because it is the source of drinking water for the one million people living in Huangshan City, but also because the river is the most important water source flowing into the Thousand Island Lake.

Located north–eastern of the Thousand Island Lake between 117°39′ and 118°54′ east longitude and 29°28′ and 30°14′ north latitude, the study watershed had an area of 5860.49 km^2 and covered the south–eastern region of Huangshan City, including the Tunxi District, She County, and part of Xiuning County. Pollutants enter the Xin'anjiang River from both point and non–point sources either directly through its vast catchment area or indirectly through its tributaries. Due to a sub–tropical, wet monsoon climate and 236 frost–free days per year, the mean annual temperature is approximately 16°C and the average precipitation is 1700 mm, with the majority of the precipitation occurring between May and August. In this study, the water quality data were obtained from eight sites along the Xin'anjiang River (Fig. 1), including Huang Shan Lin Xiao (Site 1), Heng Jiang Da Qiao (Site 2), Huang Kou Du (Site 3), Huang Dun Du (Site 4), Yu Liang (Site 5), Pu Kou (Site 6), Nan Yuan Kou (Site 7) and Jie Kou (Site 8). The entire watershed was delineated into eight sub–basins through the use of Digital Elevation Model (DEM) data (http://datamirror.csdb.cn) and readily available GIS software, in order to test for relationships between the landscape and surface water.

2.2 Data preparation

The Huangshan Environmental Monitoring Center (HSEMC) collected water samples monthly over three years (2008 to 2010). Ten common parameters were selected for this study based on sampling continuity at the eight monitoring sites, including temperature (Temp,°C), pH, electrical conductivity (EC, $ms \cdot s^{-1}$), dissolved oxygen (DO, $mg \cdot L^{-1}$), permanganate index (COD_{Mn}, $mg \cdot L^{-1}$), ammonia–nitrogen ($NH_4^+ - N$, $mg \cdot L^{-1}$), chemical oxygen demand (COD_{Cr}, $mg \cdot L^{-1}$), total nitrogen (TN, $mg \cdot L^{-1}$), total phosphorus (TP, $mg \cdot L^{-1}$) and *Escherichia coli* (*E. coli*, $num \cdot L^{-1}$). The water samples were sampled, preserved, transported, and analyzed according to the national standards for surface waters in China (GB3838–2002) in the laboratory of HSEMC. Appendix A summarizes the abbreviations, units, and analytical methods for the ten water quality parameters. The verity of the analytical processes for all parameters was ensured through careful standardization, procedural blank measurements, and spiked and duplicated samples. Table 1 compiles the basic statistics on river water quality gathered from the three year data sets.

The Multi–source Land Cover Data obtained from the Data Center for Resources and Environmental Sciences at the Chinese Academy of Sciences (http://www.geodata.cn/Portal/metadata/viewMetadata.jsp) were reclassified into the following categories: agriculture, forest, grassland, water, and urban. The data were converted from a raster set to a vector set, and then were intersected with the eight sub–basin boundaries previously derived from GIS tools. Figure 2(a) shows that forest coverage dominated nearly all of the watersheds (>60%), with the exception of Sub–basin 6, in which the percentage of agriculture coverage (45.56%) approximately equaled the percentage of forest coverage (45.48%). GIS tools were also used to calculate the slope (°), precipitation (mm), air temperature (°C), and population density ($persons \cdot km^{-2}$) of each sub–basin, based on various data sets obtained from the same land cover source. The average flow volume in each sub–basin between 2008 and 2010 was calculated based on stream–

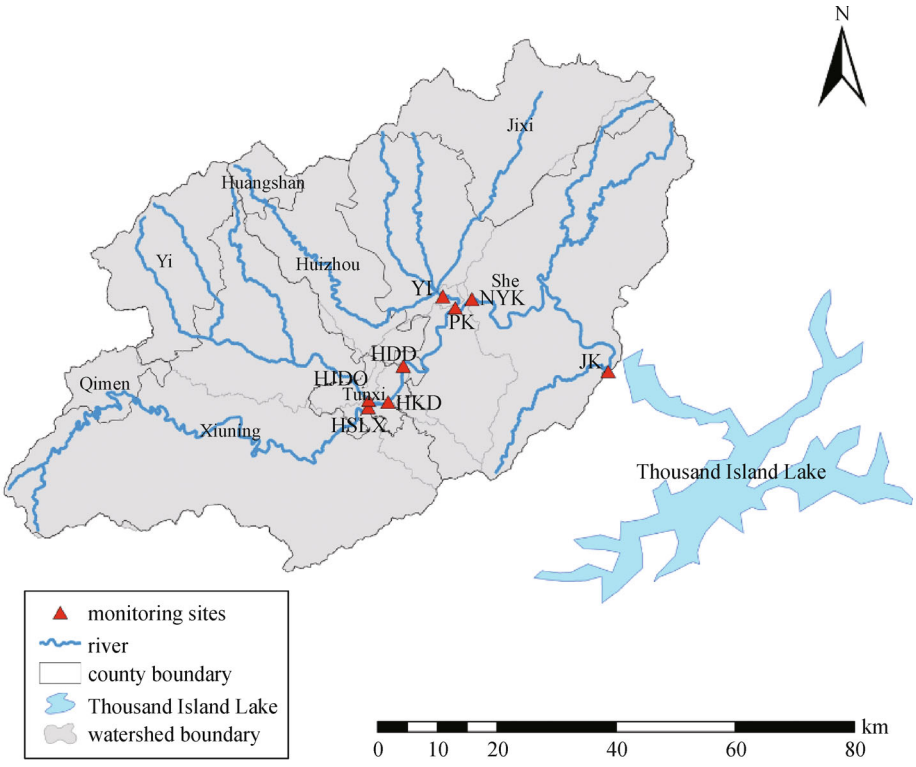

Fig. 1 Study area and eight monitoring sites

flow information obtained from the Huang Shan Hydrological Bureau. Figure 2 (b) shows how the data were standardized before being tested by statistical methods. The potential sources of pollution for each sub–basin could be deduced from studying these data side–by–side and analyzing statistical results. In addition, the temporal variations specific to the Xin'anjiang River could be explained by daily climatic data recording the precipitation and air temperature of the entire watershed.

2.3 Multivariate statistical methods

Multivariate statistical analysis of the river water quality data sets in this study was performed through a correlation matrix, CA, DA and PCA [16–18]. The mathematical tools were all applied with the following objectives: 1) to determine which clusters of months account for the variability in water quality parameters and other basin characteristics; 2) to group sampling sites by similar water pollution patterns; 3) to confirm the group results; and 4) to identify the potential sources of pollution in the Xin'anjiang River Basin.

The Spearman R coefficient, computed over ranked data, was used to account for the non–normal distribution of measured water quality parameters and basin characteristics. The Spearman R coefficient is a non–parametric measure of the correlation between variables. It is defined similarly to the Pearson correlation coefficient, but has been adapted for variables with non–normal distribution and has been computed over ranks, (i.e., the values of the variables are ranked from smallest to largest) [16,19].

CA is an unsupervised pattern recognition technique that uncovers intrinsic structures in order to group objects into clusters, which once grouped, should exhibit internal (within cluster) homogeneity and external (between clusters) heterogeneity [20,21] based on their proximity or similarity [9]. Hierarchical agglomerative cluster analysis is the most common approach that intuitively provides similar relationships between each sample and the entire data set [10]. In this study, hierarchical agglomerative cluster analysis was performed on the standardized data set by Ward's Method, using squared Euclidean distances to measure similarity. The objective of standardization is to optimize the influence of variables and to eliminate the influence of different units of measurement, which would render the data dimensionless. The result is illustrated by a dendogram, in which the linkage distance is reported as $D_{\text{link}}/D_{\text{max}}$, and represents the quotient between the linkage distance and the maximal distance, multiplied by 100 as a way to standardize the linkage distance represented on the y–axis [16,22–24].

DA is a method of analyzing dependence that occurs as a unique result of canonical correlation. One of the objectives of DA is to confirm the groups found by CA. DA constructs a discriminant function (DF) for each group, found by CA < [20,25] as follows:

Table 1 Water quality parameters with range, mean value, standard error of mean and standard deviation of Xin'anjiang River system

parameters		HSLX	HJDQ	HKD	HDD	YL	PK	NYK	JK
Temp.	range	5.4–31.5	5.6–32.5	5.5–34	5.6–34.2	4.4–29.9	5.3–30.1	3.5–31	4.2–32
	mean	17.67	17.92	18.33	18.59	18.39	18.53	18.64	19.74
	SE	1.31	1.33	1.36	1.34	1.30	1.27	1.27	1.30
	SD	7.83	7.96	8.15	8.01	7.82	7.61	7.62	7.79
pH	range	7.13–8.51	7.1–8.88	6.86–8.92	7.08–8.7	6.88–8.47	6.89–8.51	6.82–8.78	7.05–8.88
	mean	7.74	7.72	7.69	7.74	7.70	7.73	7.67	7.68
	SE	0.06	0.07	0.08	0.07	0.06	0.06	0.06	0.07
	SD	0.37	0.41	0.48	0.39	0.35	0.34	0.38	0.39
EC	range	3.1–18.6	4.2–37.7	4.2–22	4.52–22.5	15.1–69.5	13–64.2	8.6–31.8	7.8–24.9
	mean	7.02	22.78	12.13	12.81	32.42	29.20	17.18	13.76
	SE	0.49	1.39	0.81	0.71	2.40	2.18	0.95	0.60
	SD	2.93	8.37	4.85	4.24	14.40	13.08	5.71	3.59
DO	range	7.0–14.0	6.1–14.6	7.2–14.4	7.3–16.3	5.5–15.2	6.1–15.5	5.7–13.7	6.1–80.2
	mean	9.35	9.51	9.38	10.02	9.06	9.69	9.24	11.10
	SE	0.28	0.33	0.34	0.36	0.32	0.37	0.31	2.00
	SD	1.68	1.99	2.02	2.17	1.94	2.23	1.83	11.97
COD_{Mn}	range	0.6–3.8	1–5.8	1.3–3.8	1.0–3.0	1.7–5.7	1.5–4.2	1–3.8	1–3.2
	mean	1.59	2.34	2.16	2.06	3.31	2.87	2.33	1.88
	SE	0.10	0.15	0.10	0.09	0.17	0.11	0.11	0.09
	SD	0.59	0.90	0.57	0.53	1.00	0.64	0.67	0.52
NH_4^+-N	range	0.066–0.659	0.077–0.863	0.049–0.832	0.071–0.927	0.148–0.992	0.142–0.902	0.048–0.806	0.064–0.704
	mean	0.24	0.30	0.31	0.36	0.53	0.43	0.30	0.18
	SE	0.03	0.03	0.04	0.04	0.05	0.04	0.04	0.02
	SD	0.15	0.18	0.22	0.24	0.28	0.26	0.21	0.12
COD_{Cr}	range	0–9	0–17	2.2–12	0–18	5.8–19	5.0–16.0	0–16	0–10
	mean	4.31	7.92	7.38	7.38	11.16	9.88	7.93	6.06
	SE	0.40	0.48	0.34	0.57	0.56	0.40	0.50	0.41
	SD	2.39	2.86	2.04	3.42	3.38	2.39	3.00	2.49
TN	range	0.35–1.9	0.48–2.52	0.56–1.86	0.9–2.12	0.94–4.58	0.98–4.34	0.95–2.44	0.082–1.82
	mean	0.96	1.21	1.17	1.44	2.50	2.26	1.61	1.28
	SE	0.06	0.07	0.05	0.06	0.14	0.13	0.07	0.05
	SD	0.34	0.40	0.32	0.36	0.85	0.79	0.42	0.31
TP	range	0.02–0.15	0.03–0.18	0–0.14	0.03–0.14	0.04–0.2	0.04–0.18	0.02–0.16	0.008–0.09
	mean	0.06	0.07	0.06	0.08	0.11	0.08	0.07	0.04
	SE	0.00	0.01	0.00	0.00	0.01	0.01	0.00	0.00
	SD	0.03	0.03	0.03	0.03	0.04	0.04	0.03	0.02
E.coli	range	170–2400	220–3500	430–5400	430–3500	330–5400	490–5400	230–5400	220–3500
	mean	599.44	835.56	1174.08	1025.42	1523.06	1523.33	1121.11	832.50
	SE	93.36	139.51	164.46	115.06	191.65	197.10	207.12	148.25
	SD	560.17	837.03	986.75	690.38	1149.93	1182.61	1242.74	889.51

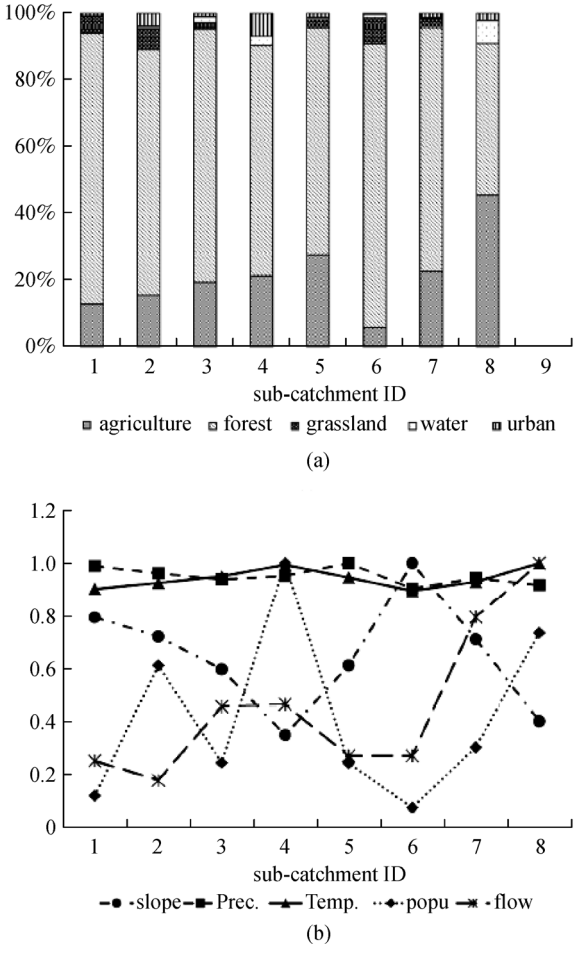

Fig. 2 (a) landscape characteristic gradients, (b) trends of temperature, precipitation, slope, population density and flow

$$f(G_i) = k_i + \sum_{j=1}^{n} w_{ij} \times p_{ij}, \quad (1)$$

where i is the number of groups (G), k_i is a constant that is inherent to each group, n is the amount of parameters used to classify a set of data into a given group, $j = 1,2,\ldots,n$, and w_{ij} is the weight coefficient, assigned by DA analysis to a given parameter (p_{ij}). In this study, DA was performed on each raw data matrix, by using standard, forward stepwise and backward stepwise modes to construct DFs to evaluate the spatiotemporal variations in the quality of river water in the basin. In this analysis, the grouping variables were the monitoring sites (spatial) and the periods (temporal), and the independent variables were all of the measured parameters. The standard DA mode constructed DFs containing all of the parameters. The forward stepwise mode added variables step–by–step, beginning with the most significant variable and continuing until no significant changes are observed. In contrast, the backward stepwise mode removed variables step–by–step, beginning with the least significant variable and continuing until no significant changes were observed [17,26].

PCA is a procedure that uses an orthogonal transformation to convert a set of potentially correlated variables into a set of linearly, uncorrelated variables, called principal components (PCs) [27,28]. PCA extracts eigenvalues and eigenvectors from the covariance matrix, which describe the dispersion of the original variables (measured parameters) in order to assess associations between variables [29]. PCA provides information about the most meaningful parameters, which describe the entire data set and thereby allow the reduction of data with minimal loss of original information [30,31].

ArcGIS software was used to obtain land use data and other basin parameters, including slope, precipitation, air temperature and population density. Microsoft Office Excel 2003 and STATISTICA 7.0 were used to make the mathematical and statistical computations.

3 Results and discussion

3.1 Temporal variations in river water quality

Temporal CA was used as an exploratory method to generate a dendrogram that grouped the 12 months into three clusters at $(D_{\text{link}}/D_{\max}) \times 100 < 20$, according to the flow volume and seasonal features, such as air temperature. As Fig. 3 shows, there were significant differences between the clusters. Empirically, 12 months would be divided into four seasons, as spring (March to May), summer (June to August), autumn (September to November), and winter (December to February). With the exception of a few discrepancies, our clusters were mostly consistent with the empirical divisions.

The seasons in the Xin'anjiang River Basin could not be classified empirically, because summer and autumn had

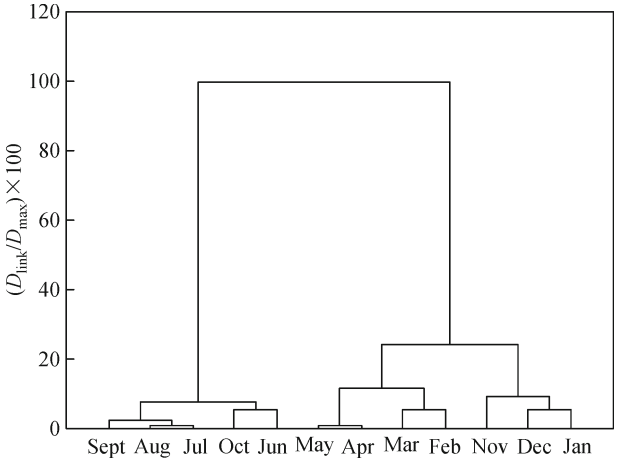

Fig. 3 Dendrogram showing temporal similarities of monitoring periods

similar temperatures and precipitation. Cluster 1 (Period 1), which included November, December and January, closely corresponded to the lowest flow period and had an average flow volume of 21.54 $m^3 \cdot s^{-1}$ and the lowest air temperature at 8.10°C. Cluster 2 (Period 2) included February, March, April and May, and Cluster 3 (Period 3) contained all the remaining months from June to October. The flow volumes were approximately the same in Periods 2 and 3, but the air temperature was significantly higher in Period 3 (25.80°C) than in Period 2 (14.52°C).

DA was applied to the raw data after dividing the entire data set into three groups, in order to avoid mistakes and to evaluate the results of temporal CA. The DA aimed to test the significance of DF and to determine the most significant variables associated with cluster differences [10]. Table shows DFs and classification matrices (CMs) obtained from the standard stepwise, forward stepwise and backward stepwise modes of DA. The DFs obtained from the standard and forward stepwise modes using ten and eight discriminant variables, yielded corresponding CMs that correctly assigned 76.39% and 77.78% of the cases, respectively. However, the DA gave CMs with 75.69% correct assignations in backward stepwise mode using only three discriminant parameters (Table 2). The temporal DA results suggested that temperature, pH and *E.coli* were the most significant parameters to discriminate between the three groups, and therefore, that the majority of expected temporal variations in the water quality were the result of these three parameters.

The mean value and standard deviation of the three discriminant parameters were respectively calculated to identify seasonal trends during the three periods (Temp: 13.05 ± 4.84, 12.78 ± 4.89, 26.28 ± 2.63; pH: 7.80 ± 0.36, 7.70 ± 0.29, 7.66 ± 0.45; *E.coli*: 1571.46 ± 1589.84, 872.88 ± 505.21, 949.17 ± 709.99). Periods 1 and 2 had similar temperatures, but the highest average Temp occurred during Period 3. The pH was relatively consistent during the three periods. High temperature may cause the decomposition of dissolved organic matter, a process that consumes great quantities of oxygen and leads to the formation of ammonia and organic acids [16,22]. The hydrolysis of these acidic materials decreases water pH values. Therefore, the pH was slightly lower in Period 2 and Period 3 than in Period 1. More *E.coli* was present in Period 1 than in the other two periods. To better understand the differences between the three groups, the analyzed variables were correlated pair–by–pair with basin characteristics, such as precipitation, temperature, and flow (Appendix B). The temperature positively correlated to the air temperature, which was consistent with the seasons. There was a significant positive correlation between precipitation and flow, but *E.coli* negatively correlated with both precipitation and flow. These results indicated that precipitation dominated the stream flow of the Xin'anjiang River, and that *E.coli* was closely related to municipal sewage and wastewater treatment plants [25,31], with constant amount and reflected the dilution effect in Period 2 and Period 3.

3.2 Spatial variations in river water quality

Just like temporal CA, spatial CA produced a dendrogram, which grouped the eight sites into three clusters at (D_{link}/D_{max}) × 100 < 16 (Fig. 4). Sites affected by similar sources were classified into groups. Site 1 and Site 8 formed Group 1; these two sub–basins had little impact from humans, relatively low pollution, and were dominated by forested areas (81.37% of Site 1 and 85.13% of Site 8). Site 5 and Site 6 formed Group 3, and were located near several water treatment plants and factories in a highly polluted region. The rest of the sites located near urban areas (e.g. Huangshan City) formed Group 2, and were

Table 2 Classification functions for discriminant analysis of temporal variation

parameters	standard			forward			backward		
	period 1	period 2	period 3	period 1	period 2	period 3	period 1	period 2	period 3
Temp.	−0.06	−0.10	0.71	−0.09	−0.13	0.69	−0.32	−0.32	0.51
pH	59.92	58.55	56.24	57.52	56.21	53.97	56.59	55.46	53.26
EC	−0.17	−0.22	−0.22	0.04	−0.02	−0.03			
DO	0.43	0.40	0.28	0.46	0.43	0.31			
COD_{Mn}	−6.12	−4.78	−4.53	−4.34	−3.05	−2.84			
NH_4^+-N	9.20	8.57	7.24	17.58	16.74	15.21			
COD_{Cr}	0.079	0.023	0.063	0.092	0.036	0.075			
TN	9.59	9.35	9.11						
TP	−0.17	−0.17	−0.16						
E.coli	0.006	0.005	0.005	0.007	0.006	0.006	0.007	0.006	0.006
constant	−244.53	−232.29	−229.38	−233.19	−221.49	−219.17	−225.57	−215.45	−214.49
%correct	53.62	73.33	99.15	53.62	71.11	99.14	34.72	79.17	98.33

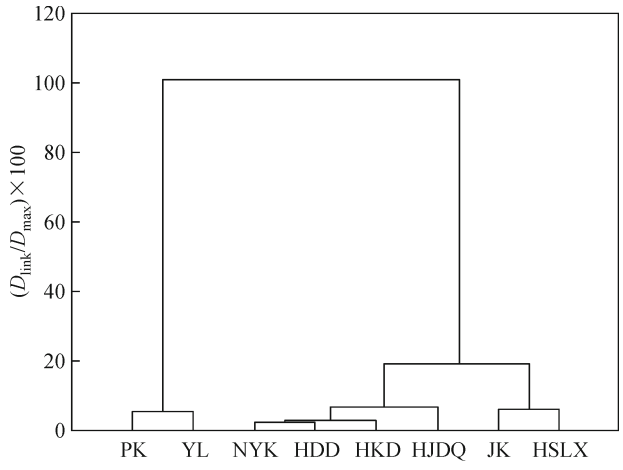

Fig. 4 Dendrogram showing sampling site clusters

moderately polluted and mainly affected by domestic wastewater.

Spatial DA was performed on the same raw data set comprised of ten parameters, after using CA to group the data into three major classes. Just as with temporal DA, DFs and CMs were obtained from the standard, forward stepwise, and backward stepwise modes (Table 3). The standard and forward stepwise mode DFs used ten and eight discriminant variables, and correctly assigned 80.56% and 79.86% of the cases to the three groups, respectively. In the backward stepwise mode, the DA produced a CM with nearly 80.56% correct assignment using only five discriminant parameters, and thereby illustrated that Temp, EC, COD_{Cr}, TN and TP were significant parameters of spatial variables (Table 3). The correct assignations through DA for these three site clusters further confirmed the adequacy of the previous spatial CA in this study, as both DA and CA verified significant differences between the three regions.

Spatial DA using backward step mode was used to calculate the mean value and standard deviation for the five selected discriminating parameters in order to evaluate different patterns associated with spatial variations in the river water quality, (Temp: 18.71 ± 7.77, 18.37 ± 7.83, 18.46 ± 7.61; EC: 10.39 ± 4.67, 16.23 ± 7.29, 30.81 ± 13.66; COD_{Cr}: 5.51 ± 1.47, 7.89 ± 2.47, 10.52 ± 2.95; TN: 1.12 ± 0.36, 1.36 ± 0.41, 2.38 ± 0.82; TP: 0.05 ± 0.03, 0.07 ± 0.03, 0.09 ± 0.04). The Spearman non–parametric correlation coefficient (Spearman R) was used in this study to establish the sub–basin characteristics associated with certain water quality parameters (Appendix C). The water temperature exhibited a closely correlated coefficient (Spearman $-R = 0.91$) to the flow. The average flow rates for Group 1, Group 2 and Group 3 were 93.52, 70.75 and 40.70 $m^3 \cdot s^{-1}$, respectively, which indicated that a higher flow volume could maintain a more stable environment but led to a higher water temperature. The trend for EC, COD_{Cr}, TN and TP suggested that Group 3 had the highest average concentration, followed by Group 2. Group 1 had the lowest average concentration. The two sub–basins in Group 3, and their watercourses, were located near a chemical plant and a sewage treatment plant, and a large amount of municipal and industrial wastewater entered this zone. Human activities significantly affected the four sites within Group 2, which were located near urban area and big counties, but Group 1 was heavily forested and had little impact from humans. Aside from human impact, basin characteristics may also correlate with water quality variables. EC, COD_{Cr}, TN and TP all significantly correlated with the percentage of agricultural area present, which was 9.13%, 33.95% and 20.69% for Group 1, Group 2 and Group 3, respectively. In addition, TN positively correlated with precipitation, while TP negatively correlated with the percentage of forest and

Table 3 Classification functions for discriminant analysis of spatial variation

parameters	standard			forward			backward		
	group 1	group 2	group 3	group 1	group 2	group 3	group 1	group 2	group 3
Temp.	0.68	0.71	0.85	0.52	0.54	0.70	0.54	0.58	0.75
pH	56.05	55.86	56.03						
EC	−0.17	−0.11	0.04	0.14	0.19	0.34	0.19	0.26	0.42
DO	0.28	0.25	0.25						
COD_{Mn}	−4.71	−4.36	−4.05	−0.11	0.29	0.52			
NH_4^+-N	4.30	5.06	3.72						
COD_{Cr}	0.62	0.78	0.95	0.79	0.96	1.11			
TN	11.76	11.95	15.46	5.14	5.43	8.77	5.81	6.44	9.98
TP	−0.19	−0.17	−0.19						
E.coli	0.004	0.005	0.005	0.0004	0.0007	0.0007			
constant	−225.31	−227.00	−242.44	−12.82	−15.28	−30.28	−10.64	−12.51	−26.67
%correct	47.22	91.67	66.67	45.83	93.06	66.67	26.39	91.67	69.44

grassland and positively correlated with air temperature. Taking account of these comprehensive factors, Site 8, the site closest to Thousand Island Lake, had better water quality than the other sites, whereas the upstream area, including Site 5 and Site 6, was highly polluted.

3.3 Source identification

PCA was employed on the normalized data to compare the compositional patterns of the water samples and to identify the influencing factors for each of the three spatial clusters. The PCA of the three data sets extracted four PCs with eigenvalues>1 for both Groups 1 and 2, and extracted three PCs for Group 3. These results explained 68.94%, 67.48% and 70.35% of the total variance in the respective water quality data from each set, respectively. Table 4 summarized the PCA results including the loadings, eigenvalue, and variance contribution rate of each PC and the cumulative variance contribution rate. Table 4 also highlighted the loading with significant absolute value in each PC.

Among the four PCs for the data set pertaining to the sites in Group 1, PC1 explained 22.92% of the total variance and had strong positive loadings (>0.70) on Temp, COD_{Mn} and $COD_{Cr.}$ As both COD_{Mn} and COD_{Cr} positively correlated to the percentage of agricultural area present, this factor may be a result of seasonal effects and of influences from non–point sources, such as agricultural activities. Anthropogenic factors did not significantly impact these two sites, due to low population densities and relatively little exploitation. The PC2 explained 19.57% of the total variance and had strong negative loadings on NH_4^+-N, which positively correlated to the percentage of agricultural area present and to population density, indicating that this factor represented the eventual contribution of ammonium fertilizers from agricultural areas to the stream via surface runoff and irrigation waters. The PC3 explained 15.60% of the total variance and had strong negative loadings on EC, indicating the impact of mineral components on surface water. The PC4 explained 10.85% of the total variance and had strong positive loadings on DO, which correlated with biochemical pollution.

Among the four PCs for Group 2, the PC1 explained 23.73% of the total variance, had strong negative loading on Temp, and dominantly represented seasonal effects, just like the PC1 of Group 1. The PC2 explained 19.97% of the total variance and had strong positive loadings on COD_{Mn}, which could be interpreted as organic pollutions from domestic wastewater, wastewater treatment plants, and agricultural activities. The five sites in Group 2 were primarily located in central Huangshan City, and received large amounts of anthropogenic pollution. The PC3 explained 12.94% of the total variance with strong negative loadings on pH, while the PC4 explained 10.84% of the total variance with strong positive loadings on EC. These two factors may be attributed to the physicochemical source of the variability [33] and to natural ioni group sources from stream inflow.

The two sites in Group 3, Site 5 and Site 6, received a great amount of municipal sewage and industrial wastewater from the Huizhou District and She County. The PC1 of this group, explaining 36.21% of the total variance, had strong positive loadings on DO and strong negative loadings on Temp. The inverse relationship between dissolved oxygen and water temperature is a natural result of warmer water becoming easily saturated with oxygen but having less capacity to hold dissolved oxygen [34].

Table 4 Loadings of 10 experimental variables on factor analysis parameters for three spatial clusters

parameters	Group 1				Group 2				Group 3		
	F1	F2	F3	F4	F1	F2	F3	F4	F1	F2	F3
Temp.	**0.71**	0.34	0.09	−0.21	**−0.89**	0.05	−0.15	−0.18	**−0.82**	−0.30	−0.16
pH	0.32	0.01	0.55	0.10	0.14	0.07	**−0.84**	0.13	0.44	0.11	**−0.70**
EC	0.16	0.12	**−0.74**	0.34	0.32	0.00	0.03	**0.82**	0.44	0.68	0.07
DO	−0.02	0.00	0.07	**0.82**	0.67	−0.36	−0.26	−0.03	**0.92**	−0.24	−0.01
COD_{Mn}	**0.89**	−0.31	0.04	0.08	−0.12	**0.81**	−0.17	−0.16	−0.05	**0.91**	0.10
NH_4^+-N	−0.06	**−0.85**	0.18	0.05	0.58	0.53	0.19	0.03	0.46	0.57	0.50
COD_{Cr}	**0.90**	0.06	−0.07	0.07	−0.01	0.86	−0.13	0.01	−0.03	**0.90**	−0.19
TN	0.17	−0.66	−0.50	0.17	0.43	0.38	0.49	0.16	0.55	0.37	0.49
TP	−0.09	−0.66	0.32	−0.40	0.52	0.19	0.13	−0.65	0.12	−0.04	0.53
E.coli	0.02	0.15	−0.63	−0.31	0.12	−0.16	0.49	0.04	0.15	0.08	**0.72**
eigenvalue	2.29	1.96	1.56	1.09	2.18	2.01	1.37	1.19	3.62	1.85	1.56
% total variance	22.92	19.57	15.60	10.85	23.73	19.97	12.94	10.84	36.21	18.51	15.63
cumulative % variance	22.92	42.49	58.09	68.94	23.73	43.70	56.64	67.48	36.21	54.72	70.35

The PC2 explained 18.51% of the total variance and had strong positive loadings on COD_{Mn} and COD_{Cr}, both of which correlated with the percentage of agricultural area present. As the COD parameter indicated the amount of organic material observable in the water, this factor represented nutrient pollution from anthropogenic sources, such as eutrophication from domestic wastewater and agricultural activities [35]. The relatively high–concentration of COD in this region made COD a significant factor for Group 1, despite the attenuating effect of the stream. The PC3, explaining 15.63% of the total variance, had strong negative loadings on pH and strong positive loadings on *E.coli*, which positively correlated with population density. This factor indicated the impact of human and animal feces; the increase of coliform led to the formation of more acidic materials, which decreased the pH values.

The major sources influencing the river water quality in all three regions were physical parameters, soluble salts, domestic wastewater, agricultural land runoff, and a small amount of industrial waste. Site 8 had better water quality than the other sites, due to the self–purification of the stream, but COD and NH_4^+-N were still significant factors on the water quality of Thousand Island Lake and should be researched further.

4 Conclusions

This case study illustrates the usefulness of multivariate statistical assessment of large and complicated databases in order to obtain meaningful information concerning the quality of surface water. Multivariate statistical methods were successfully applied to evaluate temporal and spatial variations in river water quality and to deduce the pollution sources at the monitoring sites in the Xin'anjiang River Basin. Hierarchical CA helped to group 12 months into three periods, and to divide the eight sampling sites into three groups based on their similarities regarding water–quality and natural and anthropogenic pollution sources. Discriminant analysis provided the best results for both temporal and spatial analysis. To discriminate between the seasons, DA used only three parameters (Temp, pH and *E. coli*) and had 76% correct assignations. To discriminate between the three spatial regions, DA used five parameters (Temp, EC, TN, COD_{Cr} and TP) and had 80.56% correct assignations. The results showed that most of the physical parameters followed seasonal variations, while the nutrient pollution caused significant variations in the water quality at different sites. The analysis of correlations between water quality parameters and basin characteristics showed that land cover, climate, and topography significantly influenced water quality, which should be examined further in a future study. The PCA helped to deduce the latent pollution sources for each group and to determine that the parameters responsible for water–quality variations were soluble salts (reflected by EC, COD_{Mn}, NH_4^+-N, COD_{Cr} and *E.coli*). The results showed that the multivariate statistical techniques served as excellent exploratory tools to analyze and interpret complex water quality data sets and to understand their temporal and spatial variations. This study revealed that the pollution of the Xin'anjiang River was related to both anthropogenic activities and poor wastewater management. The drinking water safety of Huangshan City and Hangzhou City depends on good water quality maintenance before the Xin'anjiang River flows into Thousand Island Lake. Therefore, the results of this study can lead to governmental consideration of strategies to mitigate further degradation and to improve the water quality in the watershed.

Acknowledgements This work was supported by the Major Science and Technology Program for Water Pollution Control and Treatment Foundation (Grant No.2008ZX07631-001) and Comprehensive Program for Water Pollution Control and Treatment (Grant No.2012A012).

Appendix is available in the online version of this article at http://dx.doi.org/10.1007/s11783-014-0736-z and is accessible for authorized users.

References

1. Liu J, Diamond J. China's environment in a globalizing world. Nature, 2005, 435(7046): 1179–1186
2. Liu J, Yang W. Water management. Water sustainability for China and beyond. Science, 2012, 337(6095): 649–650
3. Yu C, Gong P, Yin Y. China's water crisis needs more than words. Nature, 2011, 470(7334): 307–307
4. Ma X, Ortolano L. Environmental regulation in China: Institutions, enforcement, and compliance. Washington, DC: Rowman & Littlefield, 2000
5. Christophersen N, Hooper R P. Multivariate analysis of stream water chemical data: The use of principal components analysis for the end‐member mixing problem. Water Resources Research, 1992, 28(1): 99–107
6. Burns D A, McDonnell J J, Hooper R P, Peters N E, Freer J E, Kendall C, Beven K. Quantifying contributions to storm runoff through end‐member mixing analysis and hydrologic measurements at the Panola Mountain Research Watershed (Georgia, USA). Hydrological Processes, 2001, 15(10): 1903–1924
7. Huang J, Li Q, Pontius R G Jr, Klemas V, Hong H. Detecting the dynamic linkage between landscape characteristics and water quality in a subtropical coastal watershed, Southeast China. Environmental Management, 2013, 51(1): 32–44
8. Palma P, Ledo L, Soares S, Barbosa I R, Alvarenga P. Spatial and temporal variability of the water and sediments quality in the Alqueva reservoir (Guadiana Basin; southern Portugal). Science of the Total Environment, 2014, 470–471: 780–790
9. Wang Y B, Liu C W, Liao P Y, Lee J J. Spatial pattern assessment of river water quality: implications of reducing the number of monitoring stations and chemical parameters. Environmental Monitoring and Assessment, 2014, 186(3): 1781–1792

10. Mostafaei A. Application of multivariate statistical methods and water-quality index to evaluation of water quality in the Kashkan River. Environmental Management, 2014, 53(4): 865–881
11. Shi W, Zeng W.Application of *k*-means clustering to environmental risk zoning of the chemical industrial area. Frontiers of Environmental Science & Engineering, 2014, 8(1): 117–127
12. Li Q, Song J, Wei A, Zhang B. Changes in major factors affecting the ecosystem health of the Weihe River in Shaanxi Province, China. Frontiers of Environmental Science & Engineering, 2013, 7(6): 875–885
13. Wang X, Cai Q, Ye L, Qu X. Evaluation of spatial and temporal variation in stream water quality by multivariate statistical techniques: A case study of the Xiangxi River basin, China. Quaternary International, 2012, 282: 137–144
14. Wang Y, Zhang J, Feeley K, Jiang P, Ding P. Life‐history traits associated with fragmentation vulnerability of lizards in the Thousand Island Lake, China. Animal Conservation, 2009, 12(4): 329–337
15. Wang Y, Chen S, Ding P. Testing multiple assembly rule models in avian communities on islands of an inundated lake, Zhejiang Province, China. Journal of Biogeography, 2011, 38(7): 1330–1344
16. Singh K P, Malik A, Mohan D, Sinha S. Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India)—a case study. Water Research, 2004, 38(18): 3980–3992
17. Zhang Y, Guo F, Meng W, Wang X Q. Water quality assessment and source identification of Daliao River Basin using multivariate statistical methods. Environmental Monitoring and Assessment, 2009, 152(1–4): 105–121
18. Sundaray S K. Application of multivariate statistical techniques in hydrogeochemical studies—a case study: Brahmani–Koel River (India). Environmental Monitoring and Assessment, 2010, 164(1–4): 297–310
19. Alberto W D, María del Pilar D, María Valeria A, Fabiana P S, Cecilia H A, María de los Ángeles B. Marrí del Pilar D a, Marrí Valeria A, Fabiana P S, Cecilia H A, Marrí de los Ángeles B. Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. A case study: Suquía A River Basin (Córdoba–Argentina). Water Research, 2001, 35(12): 2881–2894
20. Awadallah A G, Yousry M. Identifying homogeneous water quality regions in the Nile River using multivariate statistical analysis. Water Resources Management, 2012, 26(7): 2039–2055
21. Guo L, Zhao Y, Wang P. Determination of the principal factors of river water quality through cluster analysis method and its prediction. Frontiers of Environmental Science & Engineering, 2012, 6(2): 238–245
22. Singh K P, Malik A, Sinha S. Water quality assessment and apportionment of pollution sources of Gomti river (India) using multivariate statistical techniques—a case study. Analytica Chimica Acta, 2005, 538(1): 355–374
23. Sultana J, Farooqi A, Ali U. Arsenic concentration variability, health risk assessment, and source identification using multivariate analysis in selected villages of public water system, Lahore, Pakistan. Environmental Monitoring and Assessment, 2014, 186 (2): 1241–1251
24. Gomes A I, Pires J C, Figueiredo S A, Boaventura R A. Optimization of river water quality surveys by multivariate analysis of physicochemical, bacteriological and ecotoxicological data. Water Resources Management, 2014, 28: 1345–1361
25. Heckler C E. Applied multivariate statistical analysis. Technometrics, 2005, 47(4): 517–517
26. Juahir H, Zain S M, Yusoff M K, Hanidza T I, Armi A S, Toriman M E, Mokhtar M. Spatial water quality assessment of Langat River Basin (Malaysia) using environmetric techniques. Environmental Monitoring and Assessment, 2011, 173(1–4): 625–641
27. Wang L, Wang Y, Zhang W, Xu C, An Z. Multivariate statistical techniques for evaluating and identifying the environmental significance of heavy metal contamination in sediments of the Yangtze River, China. Environmental Earth Sciences, 2014, 71(3): 1183–1193
28. Jang C S, Chen J S, Lin Y B, Liu C W. Characterizing hydrochemical properties of springs in Taiwan based on their geological origins. Environmental Monitoring and Assessment, 2012, 184(1): 63–75
29. Vega M, Pardo R, Barrado E, Debán L. Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. Water Research, 1998, 32(12): 3581–3592
30. Filik Iscen C, Emiroglu O, Ilhan S, Arslan N, Yilmaz V, Ahiska S. Application of multivariate statistical techniques in the assessment of surface water quality in Uluabat Lake, Turkey. Environmental Monitoring and Assessment, 2008, 144(1–3): 269–276
31. Li Y, Tang C, Yu Z, Acharya K. Correlations between algae and water quality: factors driving eutrophication in Lake Taihu, China. International Journal of Environmental Science and Technology, 2014, 11(1): 169–182
32. Frenzel S A, Couvillion C S. Fecal‐indicator bacteria in streams along a gradient of residential development. Journal of the American Water Resources Association, 2002, 38(1): 265–273
33. Zhou F, Huang G H, Guo H, Zhang W, Hao Z. Spatio–temporal patterns and source apportionment of coastal water pollution in eastern Hong Kong. Water Research, 2007, 41(15): 3429–3439
34. Shrestha S, Kazama F. Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan. Environmental Modelling & Software, 2007, 22(4): 464–475
35. Simeonov V, Stratis J A, Samara C, Zachariadis G, Voutsa D, Anthemidis A, Sofoniou M, Kouimtzis T. Assessment of the surface water quality in Northern Greece. Water Research, 2003, 37(17): 4119–4124