Irrigation & Drainage Systems Engineering

Research Article Open Access

Assessment of Surface Water Quality using Multivariate Statistical Techniques: A Case Study in China

Wang Y*, Zhu G and Yu R

Southeast University, Nanjing, Jiangsu, China

Abstract

In order to interpret the surface water quality of drinking water sources of Tongyu River and Mangshe River in Yancheng city, China, 18 water quality parameters were selected and data from 9 sampling sites during 2010 to 2015 from were collected and analyzed by multivariate statistical techniques, including cluster analysis (CA), principal component analysis (PCA), and factor analysis (FA). The sampling sites were classified into three clusters based on their similarities using a hierarchical CA, which represented relative low pollution sites, moderate pollution sites, and relative high pollution sites. By PCA/FA, six latent factors were identified that accounted for 75.39% of the total variance, representing the influences of organic pollution, fecal pollution, biochemical reactions, nutrients, domestic sewage, and natural factors, respectively. By pollution source analysis, the results were obtained that Sites 1, 2, and 3 were almost completely unaffected by various pollution sources, Sites 4 and 5 were polluted with industrial and domestic discharge, Sites 6, 7, and 8 were polluted with point and nonpoint sources from industrial activity, agriculture, and domestic drainage, and Site 9 was severely polluted with untreated domestic discharge from nearby inhabitants. The results verified that multivariate statistical techniques are useful, and may be necessary for analyzing and interpreting large, complex surface water quality databases, which could help managers optimize action plans to control drinking water quality.

Keywords: Water quality; Cluster analysis; Principal component analysis; Factor analysis

Introduction

Water quality has greatly deteriorated worldwide in the past decades, which is affected by both natural processes (precipitation rate, weathering processes, and soil erosion) and anthropogenic effects associated with excessive exploitation of water resources and untreated discharge of municipal and industrial wastewater [1-7]. The temporal and spatial variations in surface water quality have been monitored by governments for years in order to prevent pollution of surface water bodies. However, long-term monitoring datasets are large, with complex matrixes comprising numerous physicochemical parameters. Therefore, it is often difficult for planners to extract meaningful information from these datasets, identify significant parameters, and apportion pollution sources [7-9]. Multivariate statistical techniques such as cluster analysis (CA), principal component analysis (PCA), and factor analysis (FA) can be used to inspect complex datasets, evaluate water quality, and assess pollution sources. In recent years, a number of studies have comprehensively applied different multivariate statistical techniques in water quality assessments for optimizing monitoring networks, selecting representative water quality parameters without losing meaningful information [5,6,10-12].

In this study, 18 water quality parameters were selected and collected from 2010 to 2015 at 9 sampling stations in Yancheng city, China. The multivariate statistical methods (i.e., CA, PCA, and FA) were applied to analyze the water quality data. Firstly, similarities and dissimilarities among 9 sampling stations were classified by mean of CA. Secondly, the complex water quality datasets were analyzed to extract latent water quality factors using PCA and FA. Finally, the effects of possible pollution sources on water quality were identified.

Methods

Study area

Yancheng city (32°51′–34°12′N, 119°34′–120°27′E) is an eastern coastal district in the center of Jiangsu Province, China, with a population of more than 8 million. It is bordered by the Yellow Sea to the east, and is adjacent to Yangzhou and Huai'an cities to the west, Lianyungang city to the north, and Nantong and Taizhou cities to the south. The district covers an area of about 14,983 km^2, including 48.54 km^2 in urban districts, while the remaining area is divided into nine counties, cities, and zones including Dongtai city, Dafeng city, Xiangshui County, Binhai County, Funing County, Jianhu county, Sheyang County, Tinghu Zone, and Yandu Zone. The drinking water of Yancheng city is supplied by Mangshe River and Tongyu River, which receive pollutants from domestic sewage, agricultural runoff, aquaculture wastewater, and industrial effluent. Mangshe River originates in Dazong Lake and discharges into the East Sea; with a total length of nearly 50 km. Tongyu River has a total length of 415 km, originating from Chang Jiang River and ending in Lianyungang city. The middle reach of Tongyu River runs through Yancheng city, with a length of 183.6 km and mean flow rate of about 100 m^3/s. At Wuyou Port in Tongyu River, surface water is severely polluted by domestic wastewater from nearby settlements. Along the rivers there are nine monitoring sites (Fenghuang Bridge, Qinnanxi Bridge, Dazong Lake, Dongtai Bridge, Baiju Bridge, Shuini Bridge, Xin-Gou, Datuan Bridge, and Wuyou Bridge) (Figure 1). Water quality parameters including water temperature (T), pH, dissolved oxygen (DO), chemical oxygen demand (COD_{Cr}), 5-day biochemical oxygen demand (BOD_5), ammonia nitrogen (NH_4^+), total phosphorus (TP), total nitrogen (TN), fluoride (F^-), sulfate (SO_4^{2-}), chloride (Cl^-), nitrate (NO_3^-), alkalinity, turbidity, total dissolved solids (TDS), nitrite (NO_2^-), *Fecal* coliforms (*F. coli*), and *Escherichia* coliforms (*E. Coli*) were selected to represent

***Corresponding author:** Wang Y, School of Energy and Environment, Southeast University, Nanjing, Jiangsu, China, Tel: +86-13776415656; E-mail: wangyumin@seu.edu.cn

Received July 24, 2018; **Accepted** August 07, 2018; **Published** August 20, 2018

Citation: Wang Y, Zhu G, Yu R (2018) Assessment of Surface Water Quality using Multivariate Statistical Techniques: A Case Study in China. Irrigat Drainage Sys Eng 7: 214. doi: 10.4172/2168-9768.1000214

water quality characteristics, and analyzed semiannually from 2010 to 2015 according to standard methods (APHA, 1998). All the water quality parameters are expressed in mg L^{-1}, except temperature (°C), pH, turbidity (NTU), *fecal* coliforms (CFU/100 mL), and *Escherichia* coliforms (CFU/100 mL). The statistical summary of the water quality parameters sampled at nine monitoring site was shown in Table 1.

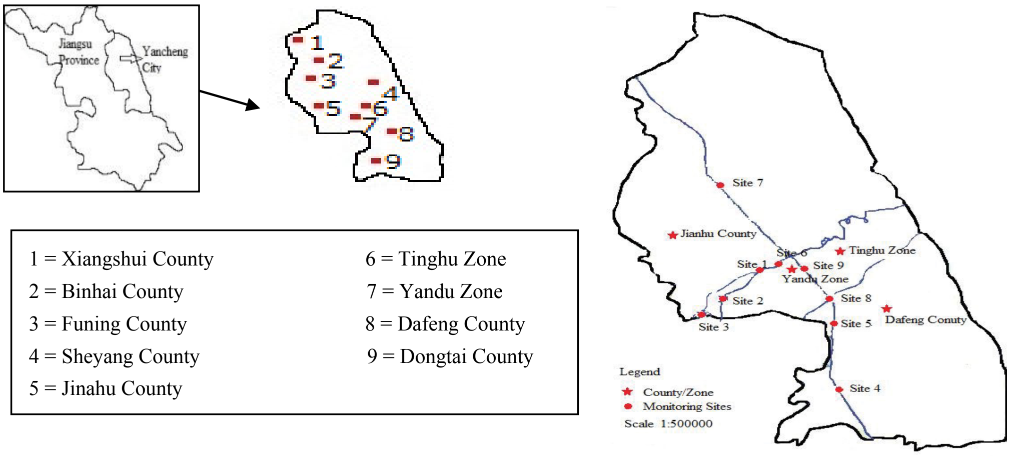

Figure 1: Sampling sites in Yancheng city.

Parameters		Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Site 7	Site 8	Site 9
T (˚C)	Range	14.3-27	14.2-26.8	14.6-26.6	5.2-27.4	5.2-27.2	16.2-19.1	14.2-19.2	5.6-26.8	18.4-26.8
	Mean	21.14	21.45	21.34	20.63	20.38	18.8	18.4	20.3	24.6
	S.D.	5.21	4.99	5.03	6.66	6.64	4.51	5.01	7.07	3.99
pH	Range	7.2-8.2	7.3-8.1	7.5-8.3	7.3-8.1	7.3-8.1	7.44-7.58	7.42-7.64	7.3-8.1	7.4-8
	Mean	7.70	7.61	7.74	7.58	7.60	7.6	7.6	7.6	7.7
	S.D.	0.29	0.26	0.33	0.19	0.23	0.12	0.09	0.26	0.32
DO (mg/L)	Range	3.3-8.5	3.5-7.7	3.1-8.6	3.5-8.1	3.90-8.60	6.5-7.1	6.8-7.1	3.1-7.2	4-5
	Mean	5.71	5.85	6.27	5.79	5.58	6.6	6.4	5.5	4.3
	S.D.	1.63	1.42	1.80	1.75	1.66	0.91	1.33	1.47	0.45
COD_{Cr} (mg/L)	Range	20.9-45.8	16.9-35.5	17.8-50.4	20.3-38.3	19.6-40.7	29.7-32.2	18.5-30.7	18.7-42.0	25.7-46.9
	Mean	33.35	29.80	33.38	31.35	30.67	29.5	28.9	32.2	33.9
	S.D.	8.21	5.06	7.88	5.22	6.24	4.82	5.88	7.54	10.21
BOD_5 (mg/L)	Range	2.0-6.0	2.0-4.0	2.0-4.0	2.0-5.0	2.0-7.0	2.5-3	2.5-4	2-3.5	1-3
	Mean	2.86	2.77	2.82	2.88	2.88	2.6	2.9	2.5	2.0
	S.D.	1.19	0.69	0.64	0.88	1.37	0.22	0.65	0.56	0.82
NH_4^+ (mg/L)	Range	0.1-1.0	0.1-0.9	0.1-1.1	0.1-0.9	0.1-1.2	0.5-1.3	0.4-0.8	0.2-1.0	0.2-0.8
	Mean	0.49	0.51	0.44	0.56	0.62	0.8	0.6	0.6	0.6
	S.D.	0.26	0.24	0.32	0.29	0.37	0.3	0.13	0.26	0.28
TP (mg/L)	Range	0.04-2.1	0.04-3.3	0.03-2.6	0.09-8.3	0.07-4.1	0.06-0.09	0.04-0.07	0.03-3	1.6-2.8
	Mean	1.03	1.01	0.80	1.82	1.55	0.1	0.1	1.4	2.4
	S.D.	0.88	1.06	0.9	2.34	1.53	0.06	0.1	1.43	0.52
TN (mg/L)	Range	0.1-3.59	0.1-4.49	0-3.07	0.1-3.62	0.1-4.09	2.98-3.18	3.03-5.12	0.2-4.11	0.2-0.8
	Mean	1.62	1.81	1.39	2.02	1.98	3.2	3.4	2.3	0.5
	S.D.	1.38	1.55	1.29	1.61	1.48	0.2	0.98	1.65	0.28
F^- (mg/L)	Range	0.4-0.77	0.41-0.74	0.45-0.7	0.34-0.59	0.46-0.68	0.49-0.6	0.49-0.82	0.4-0.7	0.5-0.6
	Mean	0.60	0.55	0.59	0.46	0.52	0.7	0.6	0.6	0.6
	S.D.	0.10	0.1	0.09	0.07	0.08	0.11	0.14	0.1	0.05
SO_4^{2+} (mg/L)	Range	26-64	29-59	29-76	33-69	27-69	45-48	49-68	36-71	23-73
	Mean	44.00	44.00	52.36	48.83	49.42	49.8	51.2	51.6	42.6
	S.D.	11.66	9.47	16.26	10.5	13.44	10.4	10.89	15.11	22.32
Cl^- (mg/L)	Range	36-87	41-69	65-129	44-112	65-122	63-97	67-103	75-118	94-134
	Mean	72.91	57.50	87.73	72.33	84.25	84.6	79.6	95.8	108.2
	S.D.	14.49	9.57	18.64	18.94	16.06	15.13	16.91	14.15	17.63
NO_3^- (mg/L)	Range	0.2-2.58	0.2-3.22	0.08-1.76	0.5-2.4	0.3-2.42	0.32-1.32	1.21-2.19	0.65-2.42	0.8-1
	Mean	0.89	1.06	0.56	1.47	1.41	1.0	1.2	1.4	0.7
	S.D.	0.69	0.78	0.48	0.64	0.66	0.82	0.68	0.6	0.38

turbidity (NTU)	Range	9-74.5	8-65	7.3-24	11-80	17-69	25.7-102.3	16.9-60.2	20-71	10-46
	Mean	34.69	33.35	13.01	56.63	52.26	42.6	51.2	48.5	33
	S.D.	16.5	17.25	5.09	48.33	37.16	33.91	27.32	15.36	15.5
Alkalinity (mg/L)	Range	137-194	111-168	112-175	102-203	0.2-195	127-207	131-193	137-189	135-176
	Mean	160.73	138.3	143.3	161.50	151.8	166.8	156.6	157.0	150.8
	S.D.	18.71	19.3	20.89	34.31	51.73	31.23	45.37	17.29	17.91
TDS (mg/L)	Range	278-458	271-476	276-497	265-479	271-481	344-451	344-456	326-484	314-456
	Mean	362.55	370.75	378.45	379.75	380.92	401.6	385.6	419.1	359
	S.D.	55.33	63.18	73.63	83.55	69.77	50.91	45.46	51.42	69.27
NO$_2^-$ (mg/L)	Range	0-0.2	0-0.12	0-0.05	0-0.2	0-0.2	0.06-0.13	0.07-0.1	0.096-0.5	0.044-0.4
	Mean	0.07	0.09	0.03	0.12	0.12	0.1	0.1	0.2	0.2
	S.D.	0.04	0.03	0.02	0.07	0.07	0.03	0.01	0.12	0.18
F. coli (CFU/100mL)	Range	100-1200	100-1000	0-450	100-2400	240-2000	260-460	120-620	140-1600	440-3200
	Mean	433	403	190	643.33	745.00	380	388.0	554.4	1155.0
	SD.	312.54	232.54	127.45	599.61	616.09	109.54	222.98	458.51	1363.56
E. coli (CFU/100mL)	Range	40-270	3-1120	3-580	73-2420	46-2420	99-152	102-914	161-1986	68-816
	Mean	149	365	136	510.89	840.22	126	508.0	659.9	357.3
	S.D.	89.16	400.56	213.56	761.96	845.52	37.48	574.17	660.54	355.69

Note: Site 1=Feng-Huang Bridge; Site 2=Qin-Xi Bridge; Site 3=Da-Zong Lake; Site 4=Dong-Tai Bridge; Site 5=Bai-Ju Bridge; Site 6=Shui-Ni Bridge; Site 7=Xin-Gou; Site 8=Da-Tuan Bridge; Site 9=Wu-You Bridge

SD: Standard deviation; DO: Dissolved oxygen; COD: Chemical oxygen demand; BOD$_5$: 5-day biochemical oxygen demand; NH4+: Ammonia Nitrogen; TP: Total Phosphorus;

TN: Total Nitrogen; F-: Fluoride; SO$_4^{2}$-: Sulfate; Cl-=Chloride; NO$_3$-; Nitrate; TDS; Total Dissolved Solids; NO$_2$-: Nitrite; *F. coli=Fecal coliforms* and *E. Coli*: *Escherichia coliforms*

Table 1: Statistical summary of the water quality parameters at nine monitoring site.

Multivariate statistical methods

Multivariate techniques including CA, PCA, and FA can reduce the dimensions of the data to enhance the quality of the analysis. Before performing CA and PCA/FA, datasets were standardized through *z*-scale transformation due to avoiding misclassification. Standardization tends to flatten the influence of variables' variance range, as well as eliminates the effects of different units among variables. All of the mathematical and statistical computations were performed using SSPS ver. 19.0 for Windows 7.

CA: CA is a multivariate technique with the primary purpose of assembling objects with respect to predetermined selection criteria, resulting in high internal (within cluster) homogeneity and high external (between clusters) heterogeneity. Hierarchical agglomerative clustering is the most common approach, which yields intuitively similar relationships between any one sample and the entire dataset, and can be represented graphically displayed as a dendrogram [3,13,14]. Dendrograms provide a visual summary of the clustering process and present a picture of the groups and their proximity with a dramatic reduction in the dimensionality of the original data [3]. Euclidean distance is usually adopted to show similarity between two samples, and can represent the difference between the analytical values from the samples [3,15].

In this study, the spatial variability of water quality was determined by hierarchical agglomerative CA on normalized datasets using Ward's method. The quotient between the links presented as Dlink/Dmax was multiplied by 100 to standardize the linkage distance [3,6,14,16].

PCA/FA: PCA is designed to form principal components (PCs), which are linear combinations of the original variables to transform the original set of inter-correlated variables into new, uncorrelated variables [6,17,18]. PCA focuses on the information from the most meaningful parameters, which minimizes the original dataset with the least loss of information [14,18]. PCA supplies an objective mode illustrating the variation in data as concisely as possible. As a result, a small number of factors can explain approximately the same amount of information as the much larger set of original observations. FA follows PCA, which further simplifies the data structure by reducing the contribution of less-significant variables by rotating the axis defined in the PCA. According to well-established rules, such as varimax rotation, new variables called varifactors (VFs) are constructed [7,14,19,20]. The difference between PCs and VFs is that PCs are a linear combination of variables (in this case, water quality variables), while VFs include unobservable, hypothetical, and latent variables [3,16,18]. In this paper, the VFs affecting river water quality were identified from large datasets using PCA/FA to distinguish possible pollution sources of sampling sites in the study area.

Results and Discussions

Spatial similarity and site grouping

Spatial CA was applied to detect similar groups among the sampling sites, and the results were presented as a dendrogram (Figure 2). All of the nine sampling sites were grouped into three statistically

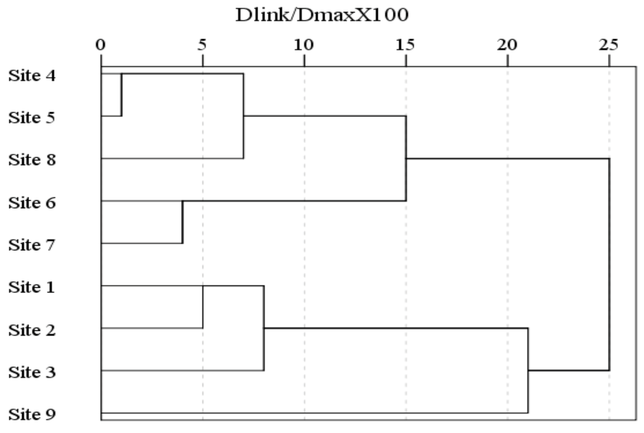

Figure 2: Dendrogram of the spatial CA

related clusters in a convincing manner using Dlink/Dmax ×.100<20. The results indicated that group A includes Sites 1, 2, and 3 located in the upstream region of Mangshe River, group B comprises Sites 4, 5, 6, 7, and 8 situated on Tongyu River and its tributary, and group C consists of Site 9, also located on Tongyu River. The three groups corresponded to relative low pollution sites (Site 1, 2, and 3), moderate pollution sites (Site 4, 5, 6 7, and 8), and relative high pollution sites, respectively. The classifications were statistically significant, because sites within the same group had similar natural and anthropogenic backgrounds. In group A, Sites 1, 2, and 3 received pollution from discharged domestic and industrial wastewater into Mangshe River. In group B, Sites 4, 5, 6, 7, and 8 were situated in the middle reaches of Tongyu River, receiving pollution from upstream sources, including domestic drainage and industrial wastewater. Finally, Site 9 (group C) received industrial pollution, domestic wastewater, and slaughter wastewater that drained into Wuyou Port, where concentrations of some water quality parameters were high, including COD_{Cr} (33.9 mg/L) and *F. coli* (1155 cfu/100 mL), while other parameters were very low such as DO (4.3 mg/L). The results indicate that hierarchical CA can provide a reliable tool to classify surface water, making it possible to design a monitoring strategy that can optimize the number of sampling sites and reduce related monitoring costs. For example, in the present study, the number of sampling sites could be reduced to one (or more) sampling site from each of groups A, B, and C to perform rapid assessments of water quality.

Data structure analysis

Correlation analysis: The correlation analysis with all the sampling stations were considered, and shown in Table 2. The results indicated that T correlates with SO_4^{2-} which is also reported in other literature [21], pH correlates with NH_4^+ since ammonia is pH-dependent [22,23], and TP has negative relationship with TN which can be explained that the river researched receive the same pollution sources [24].

Box plots of water quality parameters: The box plots of individual water quality parameters with the spatial variations corresponding to the three clusters from CA were shown in Figure 3. The water quality data of the same cluster were combined for a given parameter. The median concentration was showed by the line across the box. The first and third quartile values were showed at the bottom and top of the box. The lowest and highest observations were expressed by a vertical line extends from the bottom to the top of the box.

From Figure 3a and 3b, it can be found that group C box plots of CODcr, SO_4^{2-}, Cl^-, NO_2^-, and *F. coli* were the largest, while the smallest for DO, and TN. The reason is that group C corresponding to Site 9 is located at Wuyou bridge on Tongyu river, which receive large quantities of industrial wastewater, domestic sewage, and wastewater from pig slaughterhouse. The box plot of DO concentration in Figure 3c showed a decreasing trend in the order of group A>group B>group C, which verified that the results obtained from CA is reasonable. In box plot of CODcr concentration shown in Figure 3d, group C was the highest, while group B was the lowest, probably due to self-purification of Tongyu river. The box plots of pH and temperature showed minor difference among groups through CA. In addition, the box plots of turbidity, Alkalinity, SO_4^{2-}, F^-, and NH_4^+ also showed small differences among groups.

PCA/FA analysis: PCA is an effective pattern recognition technique used to interpret the variance of a large dataset of inter-correlated variables with a smaller set of independent components. Kaiser–Meyer–Olkin (KMO) and Bartlett's sphericity tests were performed on the parameter correlation matrix to examine the validity of the PCA. The results of the KMO and Bartlett's sphericity tests were 0.586 and 816.104, respectively, with a significance level of 0, indicating that PCA was useful for data reduction and that significant relationships were present among the variables. PCA was applied to a standardized dataset to identify the latent factors. The aim of this analysis was primarily to create an entirely new, smaller set of factors compared to the original dataset. The PCA revealed six PCs with eigenvalues>1 that explained about 75.39% of the total variance (Figure 4). These six PCs were responsible for 21.76%, 15.43%, 13.36%, 10.82%, 7.84%, and 6.18% of

Variables	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1.00																	
2	-0.29.22	1.00																
3	-0.57	0.33	1.00															
4	0.30	-0.08	-0.04	1.00														
5	-0.55.36	0.22	0.59	-0.00	1.00													
6	0.25	**-0.67**	-0.47	0.16	-0.19.08	1.00												
7	-0.20	0.17	-0.08	-0.21	-0.07	0.14	1.00											
8	0.18	-0.49	-0.06	0.03	0.03	0.20	**-0.69**	1.00										
9	0.06	0.02	-0.04	-0.01	0.06	-0.05	-0.36	0.21	1.00									
10	**-0.75**	0.24	0.30	-0.48	0.18	-0.24	0.25	-0.16	0.08	1.00								
11	0.05	-0.19	-0.15	-0.05	-0.15	0.21	-0.11	0.15	0.31	0.25	1.00							
12	-0.45	0.15	0.15	-0.38	0.11	-0.07	0.61	-0.23	-0.41	0.45	-0.16	1.00						
13	0.11	-0.05	0.10	-0.19	-0.08	-0.15	-0.06	0.17	-0.25	0.01	-0.01	0.37	1.00					
14	0.36	-0.17	-0.00	0.12	-0.08	0.01	-0.35	0.22	-0.04	-0.38	0.02	-0.26	0.12	1.00				
15	-0.43.25	-0.04	0.37	-0.28	0.16	-0.16	-0.14	0.22	0.19	0.54	0.41	0.21	0.14	0.09	1.00			
16	0.24	-0.06	-0.16	-0.23	-0.26	-0.02	0.04	0.20	-0.27	-0.06	0.11	0.1	0.42	0.12	0.07	1.00		
17	0.21	-0.28	-0.07	0.15	0.15	0.22	-0.07	0.12	-0.29	-0.24	0.07	-0.08	0.36	0.03	-0.19	0.14	1.00	
18	0.14	-0.03	0.05	-0.14.25	-0.14	-0.11	-0.05	0.14	-0.24	-0.04	-0.00	0.21	**0.65**	-0.03	-0.02	0.42	**0.58**	1.00

Note: Values in bold are corresponded to statistically significant correlation coefficients (r>0.65)(p<0.01).

1=Temprature; 2=pH: 3=DO; 4=COD_{Cr}; 5=BOD_5; 6=NH_4^+; 7=TP; 8=TN; 9=F^-; 10=SO_4^{2+}; 11=Cl^-; 12=NO_3^-; 13=Alkalinity; 14=Turbidity; 15=TDS; 16=NO_2^-; 17=*F. coli*; 18=*E. Coli*

DO: Dissolved Oxygen; COD: Chemical Oxygen Demand; BOD_5; 5-Day Biochemical Oxygen Demand; NH4+: Ammonia Nitrogen; TP; Total Phosphorus; TN: Total Nitrogen; F: Fluoride; SO_4^2-: Sulfate; Cl-: Chloride; NO3-: Nitrate; TDS: Total Dissolved Solids; NO_2-: Nitrite; *F. coli=Fecal coli form*; *E. Coli: Escherichia coliforms*

Table 2: Correlation matrix of the parameters.

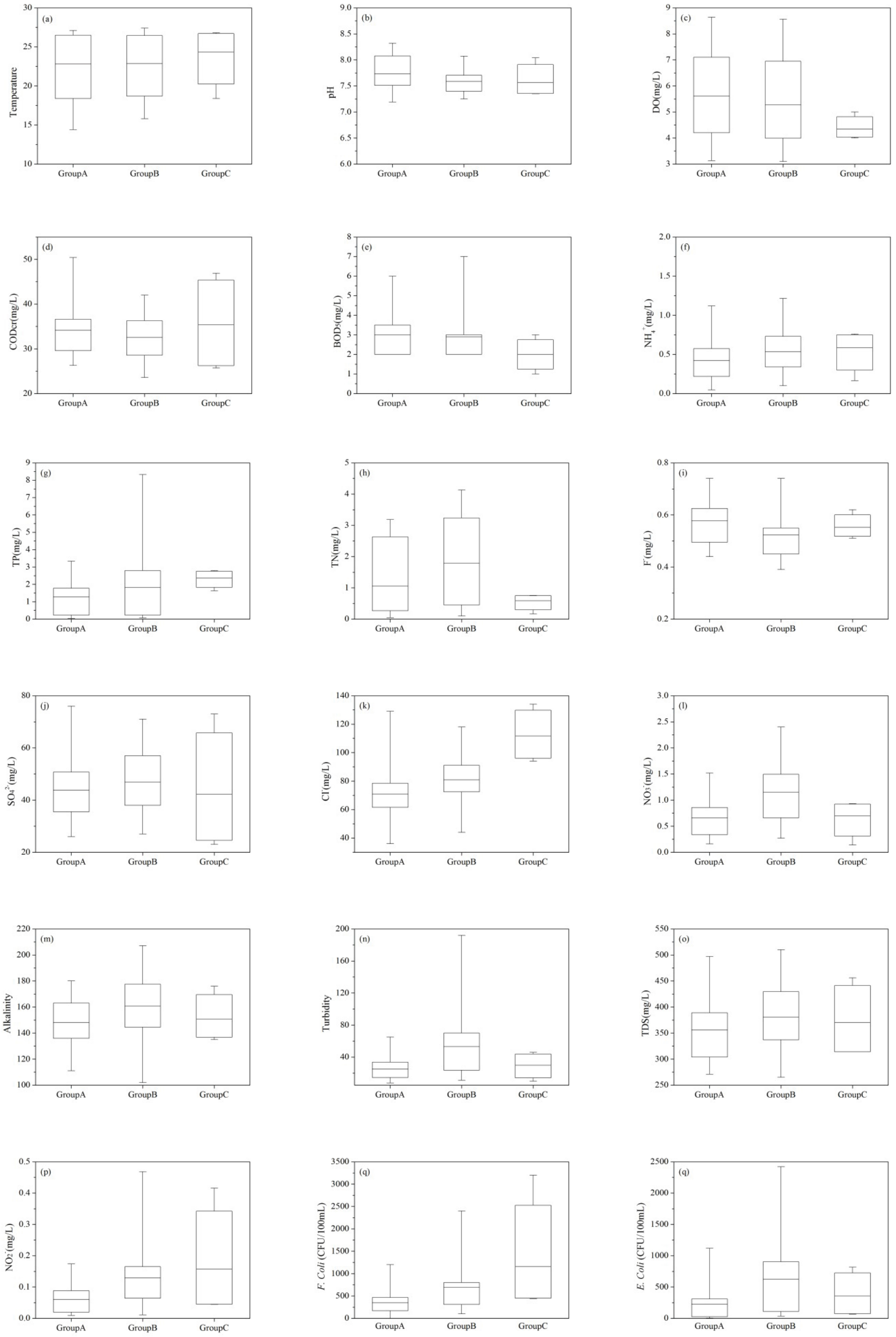

Figure 3: Box plots of 18 water quality parameters for different clusters.

the total variance, respectively. FA was performed further to reduce the contribution of less important variables to simplify the data structure resulting from the PCA. A varimax rotation of the PCs to six different VFs with eigenvalues>1 explained about 75.39% of the total variance (Table 3). As shown by the factor-loading matrix, the first VF (VF1), which explained 14.09% of the total variance, had a strong correlation with COD_{Cr} and a moderate correlation with SO_4^{2-}, nitrate, and TDS. Therefore VF1 represented organic pollution from industrial point sources. The second VF (VF2), which explained 13.69% of the total variance, was correlated heavily with turbidity, *E. coli*, and *F. coli*, representing fecal pollution from domestic point sources. The third VF (VF3), which explained 13.56% of the total variance, had positive

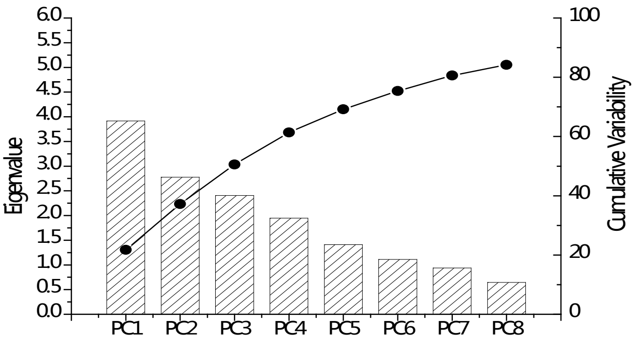

Figure 4: Scree plot of the eigenvalues of the PCs.

Element	VF1	VF2	VF3	VF4	VF5	VF6
Temp	-0.46	0.14	**-0.73**	0.34	0.07	-0.01
pH	0.03	-0.08	0.14	-0.27	**-0.86**	-0.06
DO	0.13	0.09	**0.79**	0.09	-0.33	-0.06
COD_{Cr}	**-0.73**	-0.05	0.07	0.05	0.11	0.05
BOD_5	-0.02	-0.08	**0.83**	0.04	-0.08	-0.08
NH_4^+	-0.13	-0.08	-0.22	-0.10	**0.88**	-0.01
TP	0.28	-0.09	-0.17	**-0.74**	0.04	-0.43
TN	0.04	0.19	0.10	0.68	0.40	0.29
F^-	-0.08	-0.33	0.00	0.14	-0.14	**0.73**
SO_4^{2+}	0.69	-0.05	0.34	-0.40	-0.11	0.29
Cl^-	0.19	0.09	-0.16	-0.07	0.23	**0.75**
NO_3^-	0.62	0.19	0.16	-0.39	0.05	-0.44
Turbidity	0.27	**0.77**	-0.05	0.18	-0.07	-0.15
Alkalinity	-0.04	-0.04	-0.15	**0.74**	0.00	-0.18
TDS	0.66	0.02	0.33	0.21	0.04	0.41
NO_2^-	0.36	0.47	-0.44	0.24	-0.07	-0.11
F. Coli	-0.36	**0.75**	0.01	-0.11	0.29	0.02
E. Coli	0.03	**0.91**	-0.03	-0.01	-0.07	-0.03
Eigenvalue	3.92	2.78	2.4	1.95	1.41	1.11
% of total variance	14.09	13.69	13.56	12.45	11.08	10.52
% of cumulative	14.09	27.22	41.34	53.78	64.87	75.39

Note: Values in bold are corresponded to absolute value of loading>0.70.

Table 3: Loadings of 18 parameters on significant VFs for the water quality dataset.

loading on DO and BOD_5 and negative loading on T, representing biochemical processes in the river and illustrating the fact that BOD_5 is degraded by the consumption of DO. The fourth VF (VF4) had significant loading on TP, alkalinity, and, to a lesser degree, TN, representing nutrient nonpoint sources from rainfall, agricultural runoff, atmospheric deposition, and livestock breeding. The fifth VF (VF5) had significant loading on pH and NH_4-N due to pollution from untreated nonpoint domestic discharge. The sixth VF (VF6) had positive loading on fluoride and chloride, suggestive of the effects of natural factors such as soil leaching and weathering.

Source apportionment

The main pollution sources of the rivers were urban, agricultural, industrial, and domestic wastewater. The scores of the six VFs for each sampling site are plotted in Figure 5 to show differences in the pollution sources at the sampling sites. Higher VF scores were related to factors with greater influences on a sampling site. The results indicated that pollution sources differed greatly among the sampling sites. Firstly, Sites 1, 2, and 3 (group A) had higher VF3 scores and lower VF1, VF2, VF4, VF5, and VF6 scores (except Site 3 and VF6), indicating that they were polluted mainly by organic pollutants, not markedly affected by nonpoint sources of domestic, and agricultural wastewater. Site 3 had a higher VF6 score, which illustrate that Dazong Lake receives rainfall runoff containing fluoride and chloride, as well as seepage of surrounding ground water, while the scores of VF2, VF4, and VF5 were low, which indicated that Dazong Lake was almost not affected by domestic pollution and fecal pollution. Secondly, Sites 4, 5, 6, 7, and 8 (group B) had higher VF1 and VF5 scores, indicating that they were mainly affected by organic industrial pollution and untreated domestic discharge. Among these sites, Sites 4 and 5 had similar characteristics with higher VF2 scores and moderate VF3 scores, which indicate that they also received fecal pollution from nearby pig slaughterhouse. Sites 6, 7, and 8 had higher VF4 and VF6 scores, revealing that they were polluted with agricultural drainage, livestock breeding, and nearby rainfall. Although the five sites were grouped into one cluster, there exists a great difference between their pollution sources. Finally, Site 9 (group C) was significantly affected by VF5 and moderately by VF2, illustrating that it was severely polluted with domestic drainage, also receiving industrial wastewater. From the above discussion, PCA/FA proved to be a reliable tool for distinguishing sources of pollution among sampling sites. This technique could be used to inform policies of pollution source control. In addition, it could be used to strengthen government initiatives to improve the water quality of drinking water sources.

Conclusions

In this paper, multivariate statistical techniques were applied to analyze surface water quality data from nine sampling sites in Yancheng, China. The spatial variations of surface water quality were

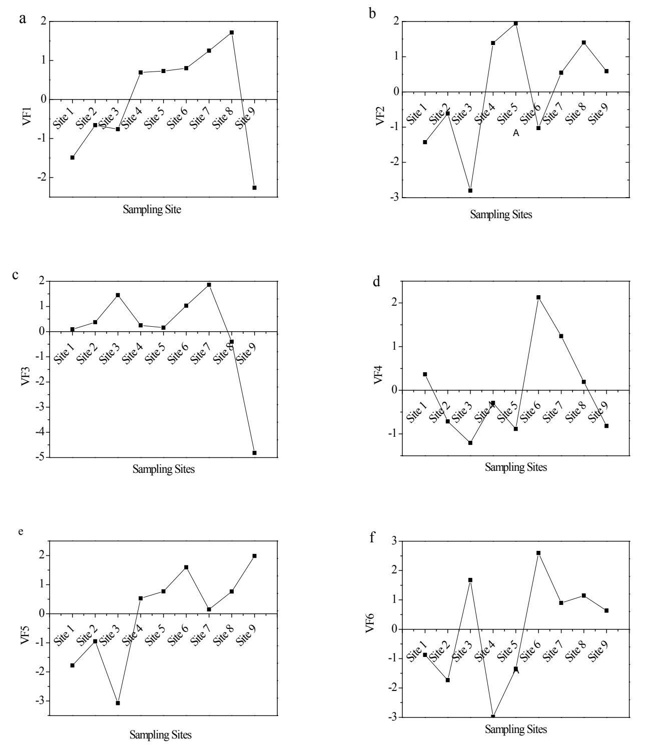

Figure 5: Scores of the five VFs at nine monitoring sites.

classified, and pollution sources of sampling sites were identified. The results obtained by hierarchical CA indicated that nine sampling sites were classified into three groups: group A (relative low pollution sites) contained Sites 1-3, group B (moderate pollution sites) included Sites 4-8, and group C (relative high pollution sites) included Site 9 solely. Through PCA/FA, six latent factors were obtained, which explained 75.39% of the total variance, and represented organic pollution, fecal pollution, biochemical reactions, nutrients, domestic sewage, and natural factors, respectively. In addition, the pollution sources of different sampling sites were analyzed according to the scores of six VFs. The results illustrated that Sites 1 and 2 were not affected greatly by pollution of nonpoint sources, Site 3 (Dazong Lake) was influenced by surrounding rain runoff, as well as ground water seepage, Sites 4 and 5 were polluted by fecal pollution, Sites 6-8 were polluted by point and nonpoint sources from industrial activity, agriculture runoff, and domestic drainage, and Site 9 was severely polluted with untreated domestic discharge from nearby residents. Based on these results, the sewage systems near Site 9 should be modified and improved by local managers as quickly as possible. The results show that multivariate statistical techniques are useful for analyzing and interpreting complex water quality datasets, as well as identifying pollution sources for governments to make effective policies. In addition, these methods can be applied by river managers to support scientific strategies to improve drinking water quality.

Acknowledgements

We thank Yancheng City Huijin Water Supply Co. Ltd for providing the water quality data. In addition, we would like to thank the editor s and referees for their valuable comments. This work was funded by the Special S & T Project on Treatment and Control of Water Pollution from the Bureau of Housing and Urban–Rural Development of Jiangsu Province (Grant No. 2014ZX07405002).

References

1. Helena B, Pardo R, Vega M, Barrado E, Fernandez JM, et al. (2000) Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga River, Spain) by principal component analysis. Water Res 34: 807-816.
2. Singh KP, Malik A, Mohan D, Sinha S (2004) Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India)—a case study. Water Res 38: 3980-3992.
3. Zhang X, Wang Q, Liu Y, Wu J, Yu M (2011) Application of multivariate statistical techniques in the assessment of water quality in the Southwest New Territories and Kowloon, Hong Kong. Environ Monit Assess 173: 17-27.
4. Chigor VN, Umoh VJ, Okuofu CA, Ameh JB, Igbinosa EO, et al. (2012) Water quality assessment: surface water sources used for drinking and irrigation in Zaria, Nigeria are a public health hazard. Environ Monit Assess 184: 3389-3400.
5. Gvozdić V, Brana J, Malatesti N, Roland D (2012) Principal component analysis of surface water quality data of the River Drava in eastern Croatia (24 year survey). J Hydroinform 14: 1051-1060.
6. Belkhiri L, Narany TS (2015) Using multivariate statistical analysis, geostatistical techniques and structural equation modeling to identify spatial variability of groundwater quality. Water Resour Manag 29: 2073-2089.
7. Chow MF, Shiah FK, Lai CC, Kuo HY, Wang KW, et al. (2016) Evaluation of surface water quality using multivariate statistical techniques: a case study of Fei-Tsui Reservoir basin, Taiwan. Environ Earth Sci 75: 6.
8. Kazi TG, Arain MB, Jamali MK, Jalbani N, Afridi HI, et al. (2009) Assessment of water quality of polluted lake using multivariate statistical techniques: A case study. Ecotoxicol Environ Saf 72: 301-309.
9. Yu C, Yin XA, Li Z, Yang Z (2015) A universal calibrated model for the evaluation of surface water and groundwater quality: model development and a case study in China. J Environ Manage 163: 20-27.
10. Simeonov V, Stratis JA, Samara C, Zachariadis G, Voutsa D, et al. (2003) Assessment of the surface water quality in Northern Greece. Water Res 37: 4119-4124.
11. Najar IA, Khan AB (2012) Assessment of water quality and identification of pollution sources of three lakes in Kashmir, India, using multivariate analysis. Environ Earth Sci 66: 2367-2378.
12. Alvarez OQ, Tagle MEV, Pascual JLG, Marín MTL, Clemente ACN, et al. (2014) Evaluation of spatial and temporal variations in marine sediments quality using multivariate statistical techniques. Environ Monit Assess 186: 6867-6878.
13. McKenna Jr JE (2003) An enhanced cluster analysis program with bootstrap significance testing for ecological community analysis. Environ Modell Softw 18: 205-220.
14. Muangthong S, Shrestha S (2015) Assessment of surface water quality using multivariate statistical techniques: case study of the Nampong River and Songkhram River, Thailand. Environ Monit Assess 187: 548-559.
15. Shrestha S, Kazama F, Newham LT (2008) A framework for estimating pollutant export coefficients from long-term in-stream water quality monitoring data. Environ Modell Softw 23: 182-194.
16. Reghunath R, Murthy TS, Raghavan BR (2002) The utility of multivariate statistical techniques in hydrogeochemical studies: an example from Karnataka, India. Water Res 36: 2437-2442.
17. Abdul-Wahab SA, Bakheit CS, Al-Alawi SM (2005) Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. Environ Modell Softw 20: 1263-1271.
18. Mohamed I, Othman F, Ibrahim AI, Alaa-Eldin ME, Yunus RM (2015) Assessment of water quality parameters using multivariate analysis for Klang River basin, Malaysia. Environ Monit Assess 187: 4182.
19. Gu H, Chi B, Li H, Jiang J, Qin W, et al. (2015) Assessment of groundwater quality and identification of contaminant sources of Liujiang basin in Qinhuangdao, North China. Environ Earth Sci 73: 6477-6493.
20. Love D, Hallbauer D, Amos A, Hranova R (2004) Factor analysis as a tool in groundwater quality management: two southern African case studies. Physics and Chemistry of the Earth, Parts A/B/C 29: 1135-1143.
21. Prathumratana L, Sthiannopkao S, Kim KW (2008) The relationship of climatic and hydrological parameters to surface water quality in the lower Mekong River. Environ Int 34: 860-866.
22. Emerson K, Russo RC, Lund RE, Thurston RV (1975) Aqueous ammonia equilibrium calculations: effect of pH and temperature. Journal of the Fisheries Board of Canada 32: 2379-2383.
23. Yerel S, Ankara H (2011) Application of multivariate statistical techniques in the assessment of water quality in Sakarya River, Turkey. Journal of the Geological Society of India 78: 1-5.
24. Chang H (2008) Spatial analysis of water quality trends in the Han River basin, South Korea. Water Res 42: 3285-3304.