Environ Monit Assess (2011) 173:17–27
DOI 10.1007/s10661-010-1366-y

Application of multivariate statistical techniques in the assessment of water quality in the Southwest New Territories and Kowloon, Hong Kong

Xuan Zhang · Qishan Wang · Yanfang Liu · Jing Wu · Miao Yu

Received: 10 February 2009 / Accepted: 11 February 2010 / Published online: 27 February 2010
© Springer Science+Business Media B.V. 2010

Abstract The application of different multivariate statistical techniques for the interpretation of a complex data matrix obtained during 2000–2007 from the watercourses in the Southwest New Territories and Kowloon, Hong Kong was presented in this study. The data set consisted of the analytical results of 23 parameters measured monthly at 16 different sampling sites. Hierarchical cluster analysis grouped the 12 months into two periods and the 16 sampling sites into three groups based on similarity in water quality characteristics. Discriminant analysis (DA) provided better results both temporally and spatially. DA also offered an important data reduction as it only used four parameters for temporal analysis, affording 84.2% correct assignations, and eight parameters for spatial analysis, affording 96.1% correct assignations. Principal component analysis/factor analysis identified four latent factors standing for organic pollution, industrial pollution, nonpoint pollution, and fecal pollution, respectively. KN1, KN4, KN5, and KN7 were greatly affected by organic pollution, industrial pollution, and nonpoint pollution. The main pollution sources of TN1 and TN2 were organic pollution and nonpoint pollution, respectively. Industrial pollution had high effect on TN3, TN4, TN5, and TN6.

Keywords Cluster analysis · Discriminant analysis · Principle component analysis · Factor analysis · Water quality

X. Zhang (✉) · Q. Wang · Y. Liu · J. Wu · M. Yu
College of Environmental Science and Engineering,
Nankai University, Tianjin, 300071,
People's Republic of China
e-mail: ahongjn@126.com

X. Zhang
School of Light Industry and Environmental Engineering, Shandong Institute of Light Industry,
Jinan, 250353, People's Republic of China

Introduction

A watercourse is a system carrying the one-way flow of a significant load of matter in dissolved and particulate phases from both natural and anthropogenic sources. The water quality at any monitoring site reflects several major influences, including the anthropogenic inputs, atmospheric inputs, climatic condition, etc. In addition, watercourses play a major role in assimilating or transporting municipal and industrial wastewater and runoff from agricultural areas. Concentrations of all kinds of pollutants have an influence on the water quality and also determine the use of water. High concentrations of toxic chemicals and biological nutrients can lead to such diverse problems as toxic algal blooms, loss of oxygen, loss of biodiversity, etc. (Ouyang et al. 2006). It is, therefore, necessary to monitor water quality,

understand the chemical and biological characteristics, and provide a reliable assessment of water quality. However, water monitoring management of a long-term period and many sampling sites produces large and complicated data sets consisting of all kinds of water parameters, which are difficult to analyze and interpret and to extract comprehensive information from them.

The application of different multivariate statistical techniques, such as cluster analysis (CA), discriminant analysis (DA), and principal component analysis/factor analysis (PCA/FA) helps in the interpretation of complicated data matrices to better understand the temporal and spatial variances of water quality, allows the identification of possible factors that influence the water systems, and offers a valuable tool for reliable management on water resources. In the last decade, comprehensive application of different multivariate statistical techniques has been gradually accepted in water quality assessment (Adams et al. 2001; Wunderlin et al. 2001; Simeonov et al. 2003; Singh et al. 2004, 2005; Kowalkowski et al. 2006; Girao et al. 2007; Kazi et al. 2008; Sojka et al. 2008).

In this study, a large data matrix, obtained from an 8-year (2000–2007) monitoring program, was subjected to different multivariate statistical techniques to extract latent information about the similarities or dissimilarities among monitoring periods and sites, to recognize water parameters responsible for temporal and spatial variances in water quality, and to identify latent factors explaining the pollution sources of sampling sites.

Monitoring area and methods

Monitoring area and sampling

The Southwestern New Territories and Kowloon area runs from the west of Tuen Mun to the east side of Kowloon and covers Western Buffer, Victoria Harbor, and North Western Water Control Zones (Fig. 1). There are five watercourses and 18 sites monitored by the Hong Kong Environmental Protection Department (HKEPD) in this area. Among them, 16 sites are monitored monthly and chosen in this paper. The watercourses and the monitored sites locating on them are Tuen Mun River (TN1, TN2, TN3, TN4, TN5, and TN6), Pai Min Kok stream (AN1 and AN2), Sam Dip Tam stream (TW1, TW2, and TW3), Kau Wa Keng (KW3), and Kai Tak Nullah (KN1, KN4, KN5, and KN7), respectively.

Monitoring parameters and analytical methods

The data for 16 water quality monitoring sites, consisting of 48 water quality parameters monthly

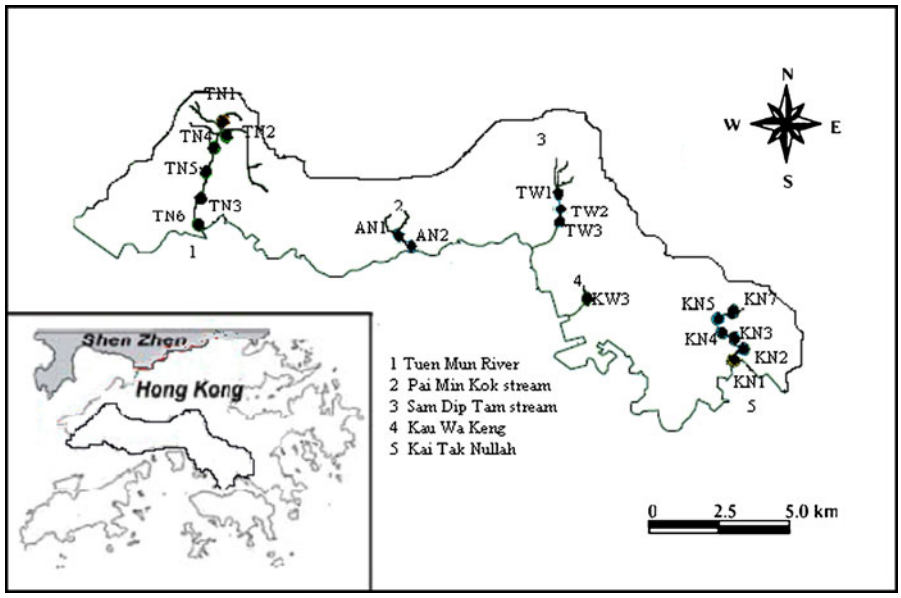

Fig. 1 Study area and sampling sites

for 8 years (2000–2007), obtained from the sampling continuity at all the selected monitoring sites, were analyzed in this study. The selected parameters included water temperature, dissolved oxygen (DO), pH, turbidity, electrical conductivity (EC), total suspended solids (TSS), total solids (TS), 5-day biochemical oxygen demand (BOD_5), chemical oxygen demand (COD), fecal coliforms (F. coli), *Escherichia* coliforms (*E. coli*), ammonia nitrogen (NH_4^+-N), nitrate nitrogen (NO_3^-–N), total Kjeldahl nitrogen (TKN), total phosphorus (TP), chromium (Cr), lead (Pb), nickel (Ni), manganese (Mn), iron (Fe), anionic surfactants (AS), fluoride(F), and zinc (Zn). Sampling and analysis for these parameters followed standard methods (APHA 1998). The basic statistics of the 8-year data set on water quality are summarized in Table 1.

Data treatment

Most multivariate statistical methods require that the data conform to normal distribution, thus, the normality distribution test of each variable was done by analyzing statistical values of kurtosis and skewness. The original data demonstrated values of kurtosis ranging from −0.198 to 1,163.709 and skewness values ranging from −0.433 to 32.367, indicating that the data were far from normal distribution. Since most of the values of kurtosis and skewness were >0, the raw data of all variables were transformed in the form $x' = \log_{10}(x)$. After the transformation, the kurtosis and skewness values ranged from −1.708 to 3.126 and −2.084 to 1.78, respectively, indicating that all the data were in normal distribution or close to normal distribution. In the case of CA, all log-transformed variables were z-scale standardized (the mean and variance were set to 0 and 1, respectively) to minimize the effects of different units and variance of variables and to render the data dimensionless.

Cluster analysis

CA is an unsupervised pattern recognition method that divides a large amount of cases into smaller groups or clusters based on the characteristics they possess. The resulting clusters of objects should exhibit high internal (within cluster) homogeneity and high external (between clusters) het-

Table 1 Statistical description of water quality parameters

Parameters	Unit	Mean	SD	Minimum	Maximum
Temperature	°C	25.00	4.11	11.8	31.8
DO	mg/L	7.16	1.82	1.20	11.90
pH		7.61	0.44	6.60	11.30
Turbidity	NTU	18.66	71.50	0.10	1,454.60
EC	µs/cm	11,623.14	12,700.26	43.00	46,410.00
TSS	mg/L	26.24	101.12	0.50	2,100.00
TS	mg/L	8,811.62	10,503.70	34.00	61,000.00
BOD_5	mg/L	8.54	13.52	0.10	210.00
COD	mg/L	20.73	26.48	2.00	540.00
F. coli	cfu/100 ml	517,598	1,669,561	1.0	31,000,000
E. coli	Cfu/100 ml	138,912	464,570	1.0	9,200,000
NH_4^+–N	mg/L	1.59	2.92	0.005	17.00
NO_3^-–N	mg/L	1.794	1.56	0.002	7.900
TKN	mg/L	2.532	4.26	0.06	65.000
TP	mg/L	0.716	1.739	0.06	61.00
Cr	µg/L	1.872	3.331	1.000	43.000
Pb	µg/L	4.634	18.572	1.00	290.00
Ni	µg/L	3.498	3.915	1.00	50.00
Mn	µg/L	142.987	72.439	10.00	2,700
Fe	µg/L	366.69	988.04	50.00	26,000
AS	mg/L	0.361	0.601	0.05	6.40
F	mg/L	0.574	0.244	0.20	2.60
Zn	µg/L	42.80	70.132	10.00	1,300.00

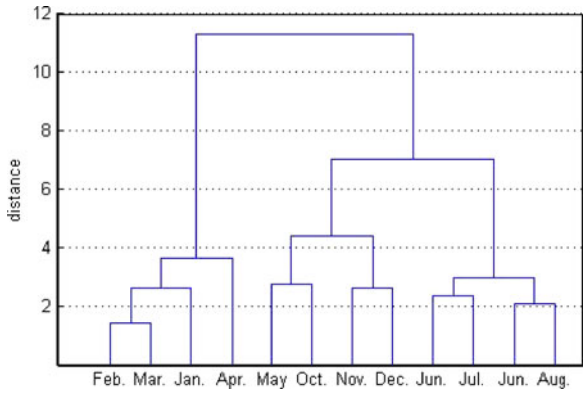

Fig. 2 Dendrogram of temporal CA

erogeneity. Hierarchical CA is the most common approach, which starts with each case in a separate cluster and joins the clusters together step by step until only one cluster remains and is typically illustrated by a dendrogram (tree diagram). The dendrogram provides a visual summary of the clustering process, presenting a picture of the groups and their proximity, with a dramatic reduction in dimensionality of original data. The Euclidean distance usually gives the similarity between two samples and a distance can be represented by the difference between analytical values from samples. In this study, hierarchical CA was performed on the standardized data by means of Ward's method, with Euclidean distance as a measure of similarity. Both temporal and spatial variances in water quality were determined from hierarchical CA using the linkage distance (Wunderlin et al. 2001; Simeonov et al. 2003; Kowalkowski et al. 2006).

Discriminant analysis

DA is used to classify cases into categorical-dependent values. One of its objectives is to determine the significance of different variables, which can allow the separation of two or more naturally occurring groups. DA operates on original data and the method constructs a discriminant function (DF) for each group as follows (Wunderlin et al. 2001; Lattin et al. 2003):

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

where i is the number of groups (G), k_i is the constant inherent to each group, n is the number of parameters used to classify a set of data into a group, and w_i is the weight coefficient assigned by DA to a given parameter (p_i).

In this study, DA was performed on original data using standard and stepwise modes to confirm the clusters determined by means of CA and evaluate both temporal and spatial variations. The best DFs were constructed considering the quality of the classification matrix (CM) and the number of parameters. The monitoring sites (spatial) and periods (temporal) were the grouping variables and all the measured parameters constituted the independent variables.

Principal component analysis/factor analysis

PCA is a powerful pattern recognition tool that attempts to explain the variance of a large data set of intercorrelated variables with a smaller set of independent variables (Simeonov et al. 2003). PCA extracts eigenvalues and eigenvectors (a list of loadings) from the covariance matrix of original variables to produce new orthogonal variables, which are linear combinations of the original variables. The principal components (PCs) provide information on the most meaningful parameters that describe a whole data set allowing data reduction with minimum loss of original information (Helena et al. 2000; Wunderlin et al. 2001).

FA follows PCA. The main purpose of FA is to reduce the contribution of less significant

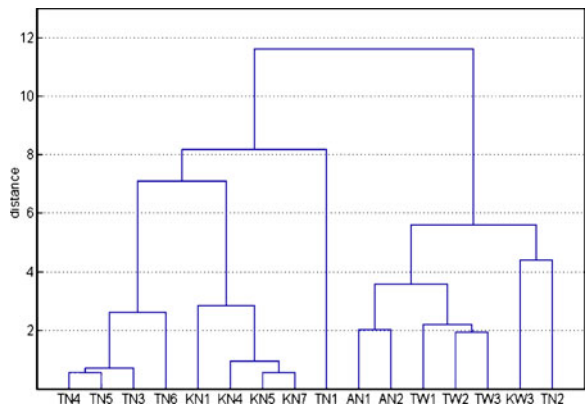

Fig. 3 Dendrogram of spatial CA

Table 2 Wilks' lambda and chi-square test of DA of temporal variation

Mode	Function	*R*	Wilks' lambda	Chi-square	*p* level
Standard mode	1	0.702	0.508	1,009.734	0.000
Stepwise mode	1	0.696	0.516	992.906	0.000

variables to simplify the data structure coming from PCA. It can be achieved by rotating the axis defined by PCA and constructing new variables known as varifactors (VFs). PC is a linear combination of observable water quality variables, whereas VF can include unobservable, hypothetical, latent variables. PCA of the normalized variables was performed to exact significant PCs and to further reduce the contribution of variables with minor significance. These PCs were subjected to varimax rotation generating VFs (Singh et al. 2004, 2005; Love et al. 2004; Abdul-Wahab et al. 2005).

Results and discussion

Temporal/spatial similarities and grouping

An initial exploratory approach involved the use of hierarchical CA on standardized log-transformed data sets sorted by the 12 months. CA generated a dendrogram (Fig. 2), grouping two clusters. Period 1 included January, February, March, and April. Period 2 included the remaining months (May, June, July, August, September, October, November, and December). The result is different from the empirical classification, which is accustomed to dividing into four seasons (spring, summer, autumn, and winter) or into dry/wet seasons. The result indicated that the temporal grouping pattern of water quality was inconsistent with the traditional classifications.

Spatial CA was used to detect similar groups between the sampling sites and produced a dendrogram (Fig. 3), grouping three clusters. Group A contained TN3, TN4, TN5, TN6, KN1, KN4, KN5, and KN7, group B comprised AN1, AN2,

Table 3 Classification functions and their coefficients for temporal DA

Parameter	Standard mode		Stepwise mode	
	Period 1	Period 2	Period 1	Period 2
Temperature	3.502	4.150	2.115	2.756
DO	2.814	2.836		
pH	51.298	51.480		
Turbidity	0.022	0.021		
EC	−0.001	−0.001		
TSS	0.019	0.016	0.002	−0.002
TS	0.001	0.001	3.44E–005	−1.5E–005
BOD_5	0.050	0.033		
COD	0.033	0.026		
F. coli	−4.02E–007	−3.77E–007		
E. coli	5.12E–006	5.41E–006		
NH_4^+–N	2.182	1.881	0.153	−0.101
NO_3^-–N	1.774	1.476	0.212	−0.161
TKN	−1.128	−1.055		
TP	1.075	1.045		
Cr	−1.417	−1.337		
Pb	−0.162	−0.164		
Ni	0.969	0.950		
Mn	0.035	0.035		
Fe	−0.005	−0.005		
AS	1.261	1.647		
F	−7.525	−8.449		
Zn	0.000	0.002		
Constant	−250.432	−265.766	−23.0732	−37.440

Table 4 Discriminant matrix for temporal DA

Mode	Period	Percent correct	Period assigned by DA	
			Period 1	Period 2
Standard mode	1	85.5	425	72
	2	84.0	161	845
	Total	84.4	586	917
Stepwise mode	1	86.1	428	69
	2	83.3	168	838
	Total	84.2	596	907

TW1, TW2, TW3, KW3, and TN2, and group C consisted of TN1. The classifications were statistically significant because the sites in these groups had similar features and natural backgrounds. Group A corresponded to relatively moderately polluted sites based on the Hong Kong Annual River Water Quality Report. Four sites (TN3–TN6) were located on the lower–middle reaches of Tuen River, receiving pollution from upstream of the Tune River and unsewered wastewater, and the other sites (KN1, KN4, KN5, and KN7) were located on the Kai Tak Nullah, which was thickly populated and received discharges from the Kowloon area. Group B corresponded to excellent water quality. In this group, except TN2, six sites were located in Pai Min Kok stream, Sam Dip Tam stream, and Kau Wa Keng, respectively, which were free from major pollution sources. The quality of these streams remained pristine over the 8 years of this study (2000–2007). Group C corresponded to highly polluted sites, which received domestic wastewater from unsewered areas. The concentration of some water parameters remained high, such as BOD_5 (44.74 mg/L), COD (65.42 mg/L), and F. coli (440,000 cfu/100 ml), and NH_4^+–N (7.43 mg/L). These results coincided with the Hong Kong Annual River Water Quality Report (HKEPD 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007).

It can be said that hierarchical CA provides a reliable tool to classify the surface water in the study area and makes it possible to optimize a future monitoring strategy which can reduce sharply the number of monitoring periods and sites and associated costs. Concretely, the monitoring frequency might be decreased and the monitoring periods could only be chosen from periods 1 and 2. Additionally, the number of monitoring sites could also be reduced to one (or more than one) sampling site from groups A, B, and C, respectively.

Temporal/spatial variations in water quality

Temporal variations in water quality parameters were evaluated by using DA with clusters based on temporal CA. The periods were the dependent variables and the measured parameters constituted the independent variables. The objectives of DA were to test the significance of DFs and to determine the most significant variables associated with the difference between the clusters. For each DF, the values of Wilks' lambda were quite small (0.508 and 0.516 for two modes, respectively) and the chi-square were rather high, indicating that the temporal DA in this study was valid and effective (Table 2).

The DFs and CMs obtained from two modes are shown in Tables 3 and 4. In the standard mode, all variables were included to construct DFs and the coefficients of *E. coli* and F. coli were close to zero. It yielded 84.4% correct assignation using 23 parameters. However, in stepwise mode, DA gave 84.2% assignation correct using only five discriminant parameters and the coefficient of TS was close to zero. The result suggested that temperature, TSS, NH_4^+–N, and NO_3^-–N were the most significant parameters for DA between two periods and accounted for most of the expected temporal variation in water quality.

The spatial DA was performed similarly to the temporal DA, as shown in Table 5. The values of Wilks' lambda and the chi-square for each DF varied from 0.045 to 0.326 and from 1,678.110 to

Table 5 Wilks' lambda and chi-square test of DA of spatial variation

Mode	Function	*R*	Wilks' lambda	Chi-square	*p* level
Standard mode	1	0.925	0.045	4,627.460	0.000
	2	0.830	0.311	1,736.713	0.000
Stepwise mode	1	0.919	0.051	4,459.623	0.000
	2	0.821	0.326	1,678.110	0.000

Table 6 Classification functions and their coefficients for spatial DA

Parameter	Standard mode			Stepwise mode		
	1	2	3	1	2	3
Temperature	2.978	2.749	3.148	2.481	2.168	2.555
DO	1.506	2.615	0.110	2.179	3.266	0.741
pH	51.853	53.949	57.024	44.331	46.164	48.875
Turbidity	0.026	0.025	0.032			
EC	−0.001	−0.002	−0.002	0.001	−7.55E–005	−2.53E–005
TSS	0.017	0.017	0.006			
TS	0.002	0.002	0.002			
BOD_5	0.133	0.092	0.229	0.104	0.216	0.080
COD	0.065	0.025	0.063			
F. coli	3.01E–009	−2.48E–007	6.67E–007			
E. coli	2.80E–006	2.22E–006	−2.93E–006	3.31E–006	2.11E–006	−7.55E–007
NH_4^+–N	3.295	1.983	3.140	1.788	1.141	2.159
NO_3^-–N	2.828	1.515	2.536	2.895	1.352	2.329
TKN	−1.429	−0.669	−0.782			
TP	1.268	0.355	0.243			
Cr	−1.475	−1.787	−1.877			
Pb	−0.142	−0.141	−0.099			
Ni	1.177	0.661	0.832	0.796	0.268	0.391
Mn	0.034	0.037	0.036			
Fe	−0.005	−0.004	−0.004			
AS	4.447	5.896	15.493	4.304	5.213	14.944
F	−5.856	−7.569	−7.014			
Zn	−0.009	0.002	−0.006			
Constant	−252.240	−255.465	−293.593	−221.473	−221.473	−254.868

4,627.460, respectively, and the p level was below 0.01, suggesting that the spatial DA was credibly effective.

DFs and CMs obtained from the standard mode and the stepwise mode of spatial DA are shown in Tables 6 and 7. The standard mode constructed DFs using 23 parameters assigning 96.3% cases correctly. However, the result of stepwise mode showed that ten parameters (temperature, pH, DO, EC, BOD_5, *E. coli*, NH_4^+–N, NO_3^-–N, Ni, and AS) were the discriminant parameters in spatial variation, with correct assignation of 96.1%. The coefficient values of EC and *E. coli* were close to zero. So it indicated that only eight parameters were needed to account for most of the expected spatial variations in water quality.

Based on the above results, the stepwise mode can produce significant DFs and recognize the most significant variables in temporal and spatial variations. It is essential to strengthen the moni-

Table 7 Discriminant matrix for spatial DA

Mode	Monitoring site	Percent correct	Assigned by spatial DA		
			Group A	Group B	Group C
Standard mode	Group A	94.5	704	33	5
	Group B	98.8	4	658	4
	Group C	90.5	1	8	86
	Total	96.3	709	699	95
Stepwise mode	Group A	94.3	700	34	8
	Group B	98.8	5	658	3
	Group C	90.5	1	8	86
	Total	96.1	706	700	97

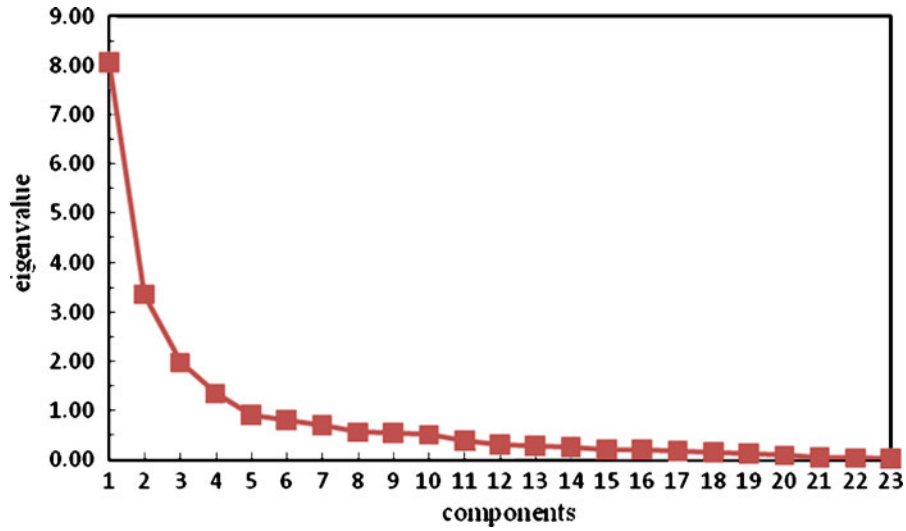

Fig. 4 Scree plot of the eigenvalues of PC

toring accuracy of the discriminant parameters for temporal and spatial DA.

Pollution sources identification in monitoring sites

Before conducting the PCA, the Kaiser–Meyer–Olkin (KMO) and Bartlett's sphericity tests were performed on the parameters correlation matrix to examine the validity of PCA. The result of KMO and Bartlett's sphericity tests were 0.845 and 33,104, respectively, indicating that PCA may be useful in data reduction.

PCA were applied to standardized log-transformed data set to identify the latent factors. The objectives of this analysis was primarily to create an entirely new set of factors much smaller in number when compared to the original data set in subsequent analysis. Based on the scree plot (Fig. 4) and the eigenvalue-one criterion, four factors were chosen as principal factors, explaining 73.24% of the total variance in the data set. The corresponding VFs, variables loadings, eigenvalues, and explained variance are presented in Table 8.

Liu et al. (2003) classified the factor loadings as "strong," "moderate," and "weak," corresponding to absolute loading values of >0.75, 0.75–0.50, and 0.50–0.30, respectively. The first factor (VF1), explaining 25.82% of total variance, had

Table 8 Loading of 23 parameters on significant VFs for water quality data set

Parameters	Four significant PCs			
	VF1	VF2	VF3	VF4
Temperature				0.71
DO			0.73	
pH				
Turbidity				
EC		0.94		
TSS				
TS		0.95		
BOD_5	0.82			
COD	0.69			
F. coli	0.55			0.69
E. coli	0.63			0.61
NH_4^+–N	0.78			
NO_3^-–N			0.86	
TKN	0.84			
TP	0.73			
Cr		0.50		
Pb				
Ni		0.74		
Mn		0.56		
Fe				
AS	0.70			
F		0.85	0.03	
Zn				
Eigenvalue	4.83	4.52	2.62	1.74
Percentage of total variance	25.82	24.15	13.98	9.29
Cumulative percentage of variance	25.82	49.97	63.95	73.24

Values above 0.5 have been shown

strong positive loadings on BOD_5, NH_4^+–N, and TKN and moderate positive loadings on COD, *E. coli*, F. coli, TP, and AS. So VF1 represented organic pollution from domestic wastewater. VF2, which explained 24.15% of total variance, had strong positive loadings on EC, TS, and F and moderate positive loadings on Cr, Ni, and Mn and represented industrial pollution. The existence of lots of ions and their compounds led to the high loadings of EC and TS. VF3, explaining 13.98% of total variance, had strong positive loading on NO_3^-–N and moderate positive loading on DO. This factor represented the contribution of non-point source pollution from agriculture areas. In these areas, farmers use the nitrogenous fertilizer and the watercourses receive nitrate nitrogen via groundwater leaching and runoff. VF4, explaining 9.29% of total variance, had moderate positive loadings on temperature, F. coli, and *E. coli* and represented fecal pollution.

To identify the pollution sources of different monitoring sites, the scores of the four VFs for each monitoring site were plotted in Fig. 5. High score corresponded to high influence of the factor on the sampling site. It indicated that the pollution source of the monitoring sites differed significantly. KN1, KN4, KN5, and KN7 had high scores of VF1 and VF2, indicating that they were greatly affected by organic pollution and industrial pollution. TN3, TN4, TN5, and TN6 had high scores for VF2 and meant that their main pollution source was industrial pollution. It can be seen that, although the four monitoring sites on Tuen Mun River (TN3–TN6) and the four monitoring sites on Kai Tak Nulluch (KN1, KN4, KN5, and KN7) were grouped into one cluster based on CA result, their pollution sources differed greatly. The same condition also existed between TN2 and the rest monitoring sites of group B (AN1, AN2, KW3, TW1, TW2, and TW3). Based on CA result,

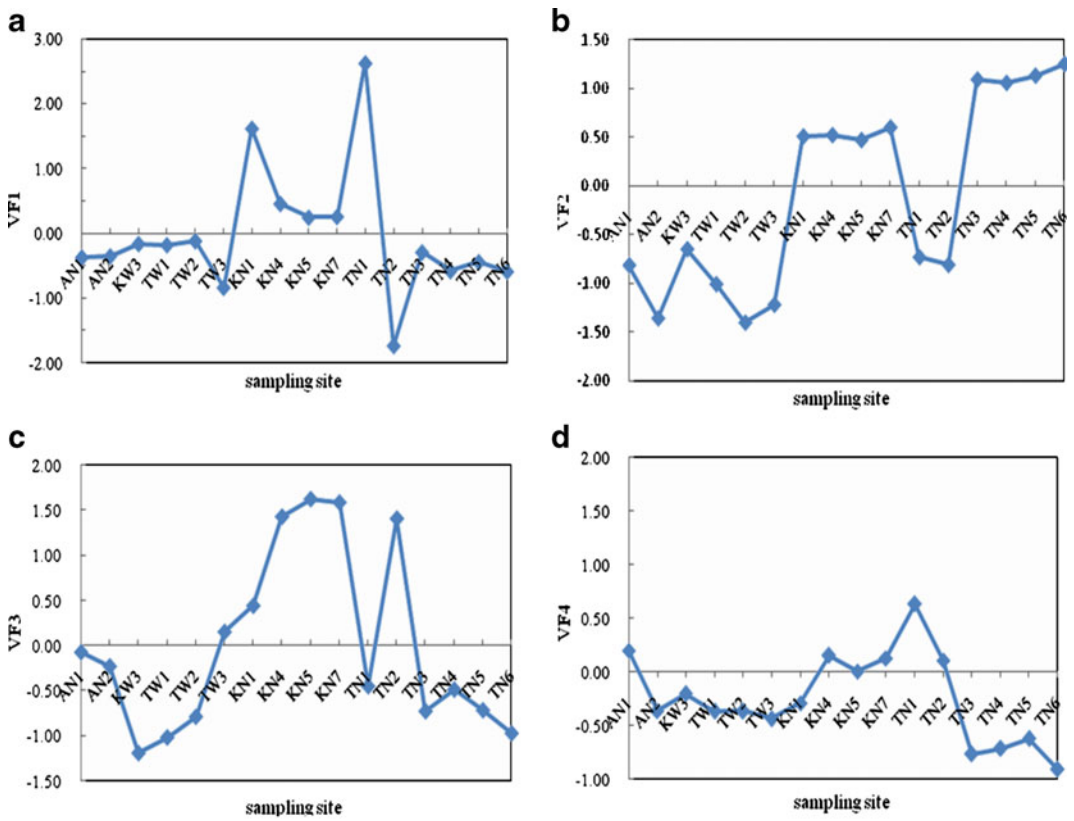

Fig. 5 Scores of the four VFs for 16 monitoring sites

AN1, AN2, KW3, TW1, TW2, TW3, and TN2 were classified into group B whose water quality was excellent. However, TN2 had a high score of VF3, which suggested that its main pollution source was a nonpoint source, while the scores of four VFs in AN1, AN2, KW3, TW1, TW2, and TW3 were low, which indicated that the effect of four VFs on them was the least. TN1 was highly influenced by organic pollution and fecal pollution, while the industrial pollution and nonpoint pollution had low effect on it.

From the above discussion, we can say that PCA/FA is a useful tool to analyze the pollution source of monitoring sites. It can offer information to identify pollution sources and help in the decision making on controlling water pollution.

Conclusions

Multivariate statistical techniques were successfully applied to evaluate temporal and spatial variations in water quality and source identification at the monitoring sites in Southwestern New Territories and Kowloon area, Hong Kong, indicating that different methods were effective and harmonious with each other and were useful for water quality management. Hierarchical CA grouped the 12 months into two periods and classified the 16 monitoring sites into three groups based on the similarity of water quality characteristics. DA provided significant DFs and recognized the most significant parameters both temporally and spatially. It yielded a dramatic data reduction, as it used only four parameters (temperature, TSS, NH_4^+–N, and NO_3^-–N), affording 84.2% correct assignations in temporal analysis, and eight parameters (temperature, pH, DO, BOD_5, NH_4^+–N, NO_3^-–N, Ni, and AS), affording 96.1% correct assignations in spatial analysis. Therefore, DA allowed a reduction in the dimensionality of the large data set. Based on the above analysis, the temporal and spatial similarity and difference may allow optimization of the monitoring frequency, the number of sampling sites, the number of parameters monitored, and thus the associated cost. PCA/FA identified four latent factors and explained 73.24% of the total variance, standing for organic pollution, industrial pollution, nonpoint pollution, and fecal pollution, respectively. KN1, KN4, KN5, and KN7 were greatly affected by organic pollution, industrial pollution, and nonpoint pollution. The main pollution source of TN1 and TN2 were organic pollution and nonpoint pollution, respectively, while industrial pollution had high effect on TN3, TN4, TN5, and TN6. Thus, this study illustrated the usefulness of multivariate statistical techniques for the analysis and interpretation of complex data set, water quality assessment, and identification of pollution sources.

Acknowledgements All thanks should first go to Hong Kong Environmental Protection Department (HKEPD) for the permission to use the data. In addition, we would like to show our great gratitude to the editors and referees for their valuable comments. The opinions in this paper are those of the authors and do not reflect the views or policies of HKEPD.

References

Abdul-Wahab, S. A., Bakheit, C. S., & Al-Alawi, S. M. (2005). Principal component and multiple regression analysis in modeling of ground-level ozone and factors affecting its concentrations. *Environmental Modelling & Software, 20*(10), 1263–1271.

Adams, S., Titus, R., Pietesen, K., Tredoux, G., & Harris, C. (2001). Hydrochemical characteristic of aquifers near Sutherland in the Western Karoo, South Africa. *Journal of Hydrology, 241*(1–2), 91–103.

APHA (1998). *Standard methods for the examination of water and wastewater*. Washington: American Public Health Association.

Girao, E. G., de Andrade, E. M., Rosa, M. D., de Araujo, L. D. P., & Meireles, A. C. M. (2007). Water quality assessment of the Jaibaras River, Ceara, Brazil using principal component analysis. *Revista Ciencia Agronomica, 38*(1), 17–24.

Helena, B., Pardo, R., Vega, M., Barrado, E., Fernández, J. M., & Fernández, L. (2000). Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga River, Spain) by principal component analysis. *Water Research, 34*(3), 807–816.

HKEPD (2000). *River water quality in Hong Kong in 2000*. Hong Kong: Hong Kong Government Printer.

HKEPD (2001). *River water quality in Hong Kong in 2001*. Hong Kong: Hong Kong Government Printer.

HKEPD (2002). *River water quality in Hong Kong in 2002*. Hong Kong: Hong Kong Government Printer.

HKEPD (2003). *River water quality in Hong Kong in 2003*. Hong Kong: Hong Kong Government Printer.

HKEPD (2004). *River water quality in Hong Kong in 2004*. Hong Kong: Hong Kong Government Printer.

HKEPD (2005). *River water quality in Hong Kong in 2005*. Hong Kong: Hong Kong Government Printer.

HKEPD (2006). *River water quality in Hong Kong in 2006*. Hong Kong: Hong Kong Government Printer.

HKEPD (2007). *River water quality in Hong Kong in 2007*. Hong Kong: Hong Kong Government Printer.

Kazi, T. G., Arain, M. B., Jamali, M. K., Jalbani, N., Afridi, H. I., Sarfraz, R. A., et al. (2008). Assessment of water quality of polluted lake using multivariate statistical techniques: A case study. *Ecotoxicology and Environmental Safety, 72*(2), 301–309.

Kowalkowski, T., Zbytniewski, R., Szpejna, J., & Buszewski, B. (2006). Application chemometrics in river water classification. *Water Research, 40*(4), 744–752.

Lattin, J., Carroll, D., & Green, P. (2003). *Analyzing multivariate data*. New York: Duxbury.

Liu, C. W., Lin, K. H., & Kuo, Y. M. (2003). Application of factor analysis in the assessment of groundwater quality in a Blackfoot disease area in Taiwan. *Science in the Total Environment, 313*(1–3), 77–89.

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, 29*(15–18), 1135–1143.

Ouyang, Y., Nkedi-Kizza, P., Wu, Q. T., Shinde, D., & Huang, C. H. (2006). Assessment of seasonal variations in surface water quality. *Water Research, 40*(20), 3800–3810.

Simeonov, V., Stratis, J. A., Samara, C., Zachariadis, G., Voutsa, D., Anthemidis, A., et al. (2003). Assessment of the surface water quality in Northern Greece. *Water Research, 37*(17), 4119–4124.

Singh, K. P., 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 Research, 38*(18), 3980–3992.

Singh, K. P., Malik, A., & Sinha, S. (2005). Water quality assessment and appointment of pollution sources of Gomti River (India) using multivariate statistical techniques: A case study. *Analytica Chimica Acta, 538*(1–2), 355–374.

Sojka, M., Siepak, M., Ziola, A., Frankowski, M., Murat-Blazejewska, S., & Siepak, J. (2008). Application of multivariate statistical techniques to evaluation of water quality in the Mala Welna River (Western Poland). *Environmental Monitoring & Assessment, 147*(1–3), 159–170.

Wunderlin, D. A., Diaz, M. P., Ame, M. V., Pesce, S. F., Hued, A. C., & Bistoni, M. A. (2001). Pattern recognition techniques for the evaluation of spatial and temporal variations on water quality. A case study: Suquira river basin (Cordoba-Argentina). *Water Research, 35*(12), 2881–2894.