J. Cent. South Univ. (2015) 22: 3770−3780
DOI: 10.1007/s11771-015-2921-z

Springer

Assessment of temporal and spatial variations in surface water quality using multivariate statistical techniques: A case study of Nenjiang River basin, China

ZHENG Li-yan(郑力燕), YU Hong-bing(于宏兵), WANG Qi-shan(王启山)

College of Environmental Science and Engineering, Nankai University, Tianjin 300071, China

Abstract: Assessment of temporal and spatial variations in surface water quality is important to evaluate the health of a watershed and make necessary management decisions to control current and future pollution of receiving water bodies. In this work, surface water quality data for 12 physical and chemical parameters collected from 10 sampling sites in the Nenjiang River basin during the years (2012−2013) were analyzed. The results show that river water quality has significant temporal and spatial variations. Hierarchical cluster analysis (HCA) grouped 12 months into three periods (LF, MF and HF) and classified 10 monitoring sites into three regions (LP, MP and HP) based on the similarity of water quality characteristics. The principle component analysis (PCA)/factor analysis (FA) was used to recognize the factors or origins responsible for temporal and spatial water quality variations. Temporal and spatial PCA/FA revealed that the Nenjiang River water chemistry was strongly affected by rock/water interaction, hydrologic processes and anthropogenic activities. This work demonstrates that the application of HCA and PCA/FA has achieved meaningful classification based on temporal and spatial criteria.

Key words: Nenjiang River basin; water quality; hierarchical cluster analysis (HCA); principal component analysis (PCA); factor analysis

1 Introduction

Water, used by households, agriculture, and industry, is clearly the most important good provided by freshwater systems. However, in an industrialized society, maintaining completely unpolluted water in all drains, streams, rivers, and lakes is probably impossible, especially in China [1−2]. Pollution of surface water with toxic chemicals and eutrophication of rivers and lakes with excess nutrients are of great environmental concern worldwide [3]. The degradation of water quality due to these contaminants has resulted in species composition alteration and overall health decrease of aquatic communities within the river basin [4]. With an increased understanding of the importance of drinking water quality to public health and raw water quality to aquatic life, there is a great need to assess water quality.

Rivers and streams are highly heterogeneous at different spatial scales [5]. This spatial heterogeneity may be controlled by complex anthropogenic and natural factors [6]. The anthropogenic discharges can be considered a constant polluting source, but not so the surface runoff which is seasonal and highly affected by climate [7]. Seasonal variation in precipitation, surface runoff, ground water flow, interception and abstraction strongly affect flow rates and consequently the concentrations of chemical compositions of river water [6−7]. In addition, pollutants entering a river system normally result from many transport pathways including storm water runoff, discharge from ditches and creeks, vadose zone leaching, groundwater seepage, and atmospheric deposition. These pathways are seasonal-dependent [3]. Therefore, due to the seasonality and regionality of river water, assessing spatial and temporal variations of river water quality at a watershed level has become an important aspect for the physical and chemical characterization of aquatic environments [3, 5−7].

In order to restore the health of the river water quality and prevent its further pollution, one of such critical efforts was the development of the surface water monitoring network [3, 8−9]. However, although such long-term survey and monitoring programs are very critical to a better knowledge of hydrology, geochemistry, and pollution in the river, they produce large sets of data that are often difficult to interpret and to draw meaningful conclusions [7, 10−11]. Further, for effective pollution control and water resources management, it is required to identify the pollution sources and the most significant parameters contributing to spatial and temporal variations.

Foundation item: Project(2012ZX07501002−001) supported by Major Science and Technology Program for Water Pollution Control and Treatment of the Ministry of Science and Technology, China
Received date: 2014−09−18; **Accepted date:** 2014−11−30

Corresponding author: YU Hong-bing, Professor; Tel: +86−22−23503797; E-mail: hongbingyunk@126.com

The problems of data reduction and interpretation, characteristic change in water quality parameters, and indicator parameter identification can be approached through the use of multivariate statistical techniques, such as cluster analysis (CA), principal component analysis (PCA), and factor analysis (FA) [3, 7, 12]. Cluster analysis (CA) has been used by many authors to study the similarities in water quality measured at differing locations and is one of the most frequently applied techniques [7, 12−18]. Principal components analysis (PCA) greatly reduces the dimensionality of the variable space by extracting a smaller number of linear combinations of the original variables, principal components. Factor analysis (FA) attempts to identify the relationships among variables, to identify representative variables from a large set of variables, and create a new (smaller) set of variables replacing the original variables for future analysis [19].

The Nenjiang River basin is an important foodstuff base and eco-environmental fragile area in China. Currently, the watershed of the Nenjiang River as the main source of surface water plays a key role in agricultural irrigation, socio-economic development, hydropower generation, wetland recharge and local eco-environmental conservation in the basin. Located in the temperate and monsoon climatic zone, the study area has a typical continental climate with long, extremely cold, and dry winter and short, mild and moist summer. The warming and drying trend of the basin has influenced seasonal stream flow (Fig. 1) [20]. In addition, rapid urbanization along the river plays an important role in the increase of point and non-point source pollution loading. Thus, the variability and quality of the water from this watershed has attracted attentions and interests from academic circles and local government. Despite its significance, there is a lack of knowledge regarding the water quality of the Nenjiang River and its temporal and spatial variations.

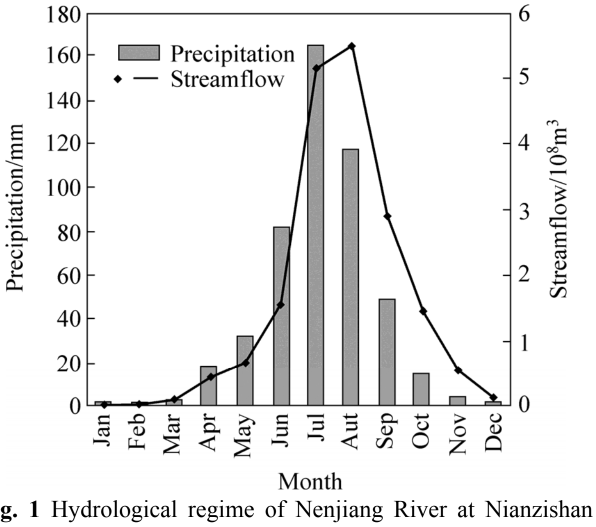

Fig. 1 Hydrological regime of Nenjiang River at Nianzishan station, based on records from 1956−2006 (Adapted from Ref. [20])

Therefore, this work attempts to apply the HCA and PCA/FA techniques to evaluating the temporal and spatial variations of water quality parameters from the viewpoint of the whole basin. The objectives of this work are to reveal the temporal and spatial variations in water quality, and identify factors and sources influencing the chemistry of the river water. The overall aim of the present work is to provide useful information for water resources management at the watershed scale.

2 Materials and methods

2.1 Study area

The Nenjiang River (45° 27′−51° 38′ N, 119°52′−126° 30′ E), the largest tributary of the Songhua River，is located in the northeast part of China, with a total length of 1370 km and a drainage area of 2.97×10^5 km^2, which includes multiple tributaries (Fig. 2). The Nenjiang River originates from Greater Khingan Mountains, wanders through the southern semiarid region, passes through the western plain, and finally discharge into the Songhua River.

The average annual temperature varies from −4.63 °C in the humid mountains to 6.43 °C in the semi-arid eastern plains. The extreme minimum and maximum temperatures in history records are −39.5 °C and 40.1 °C, respectively. The annual precipitation mainly concentrates in June to September, which accounts for 70%−80% of total precipitation [20−21]. The main land use types of the study area are forest (35.44%) and agricultural lands (30.88%). The remaining area of the basin is covered by pasture (20.08%) and wetlands (6.25%) [20].

The Nenjiang River basin can be divided into three distinct areas: the upper, middle, and lower basins on the basis of various interrelated factors such as altitude, climate, topography, physiognomy. Mainly dominated as piedmont of Greater Khingan Range and Lesser Khingan Range, the upper basin is a mountainous forest region with good vegetation and complicated topography. The middle and lower reaches of the basin are morphologically dominated by hills and plains with rich mineral resources, fertile soil, relatively dense population as well as developed industry and agriculture. The Nenjiang River receives pollution load from both the point and non-point sources. Municipal wastewater is directly discharged into the river as the wastewater and sewage treatment plant is still being built. Moreover, yet big industries settled in the area purify their wastewater, small industries are suspected to discharge residues into the river. Besides, it receives agricultural run-off from its vast catchments area directly or through its tributaries

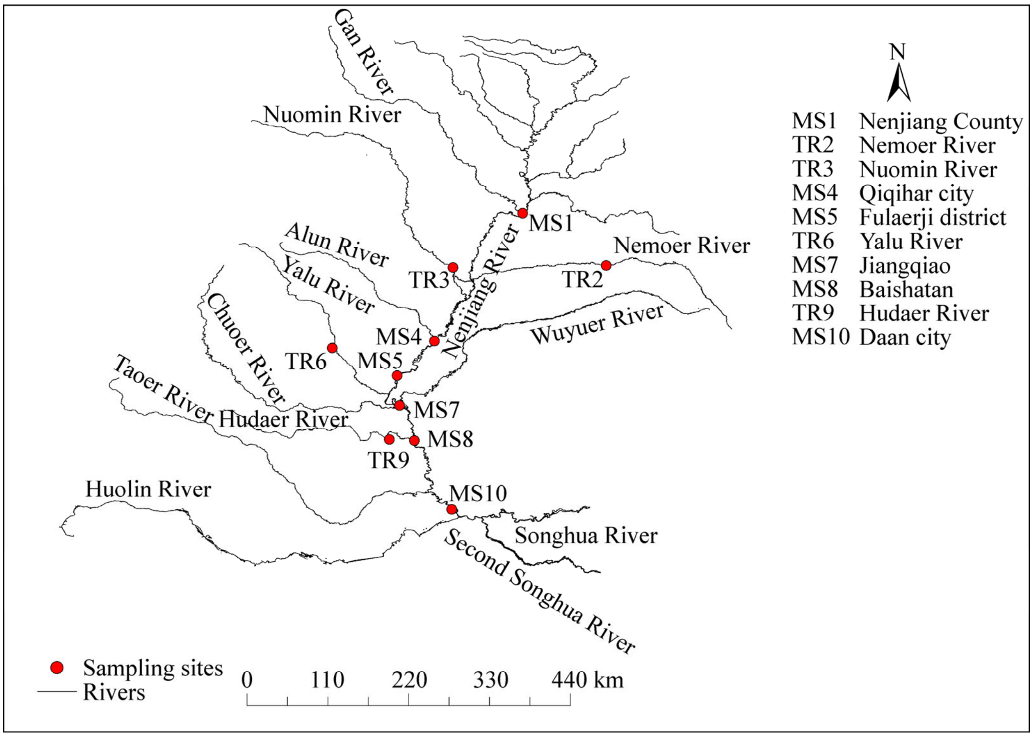

Fig. 2 Map of study area and surface water quality sampling sites in Nenjiang River basin, China

and wastewater drains. The combination of extreme continental climate causes river hydrology and hence river pollution to be strongly influenced by seasonality.

2.2 Data

Based on the hydrologic river features and on our previous results, we selected 10 sampling sites located in the Nenjiang River basin (Fig. 2). The sampling sites were designed to cover a wide range of determinants at key sites, which reasonably represent the water quality of the river system accounting for tributary and inputs from wastewater drains that have impact on downstream water quality. Selected sites were sampled every month for two years (2012−2013).

Although more than 20 water quality parameters were available, only 12 representative parameters (Table 1) were selected for testing due to their continuity in measurement at all 10 sampling sites. The selected parameters for the assessment of surface water quality characteristics included water temperature (Temp), transparency (SD), pH, dissolved oxygen (DO), electrical conductivity (EC), ammonium nitrogen (NH_3-N), total organic carbon (TOC), chemical oxygen demand ($KMnO_4$) (COD), 5-day biological oxygen demand (BOD_5), total nitrogen (TN), total phosphorus (TP), and chlorofucine α (Chla). All the sampling and pretreatment of samples followed the Chinese National Standards for Scientific Sampling (Ministry of Environmental Protection of China, 2002).

Table 1 Water quality parameters associated with their abbreviations and units used in this work

Parameter	Abbreviation	Unit
Water temperature	Temp	°C
Transparency	SD	m
pH	pH	pH
Dissolved oxygen	DO	mg/L
Electrical conductivity	EC	mg/L
Ammonia nitrogen	NH_3-N	mg/L
Total organic carbon	TOC	mg/L
Chemical oxygen demand ($KMnO_4$)	COD	mg/L
5-day biological oxygen demand	BOD_5	mg/L
Total nitrogen	TN	mg/L
Total phosphorus	TP	mg/L
Chlorofucine α	Chla	mg/L

2.3 Data treatment

Correlation structure between variables was studied using the Spearman *R* coefficient, in order to account for variables with non-normal distribution [13]. To examine the suitability of the data for PCA/FA, Kaiser-Meyer-Olkin (KMO) and Bartlett's Sphericity tests were performed [17]. HCA, PCA and FA applied to experimental data standardized through *z*-scale (mean=1, variance=0) transformation in order to avoid misclassification due to wide differences in data dimensionality. Standardization tended to minimize the effects of differences in measurement units and variance of variables and rendered the data dimensionless [7, 12, 22−23].

2.4 Cluster analysis

Cluster analysis (CA) classifies objects (cases) into classes (clusters/groups) so that each object is analogous to the others in the cluster but different from those in other classes [13, 19]. The results of CA help in interpreting the data set and indicating the patterns [7]. Hierarchical cluster analysis (HCA) is the most common approach, which starts with the most similar pair of objects and forms higher clusters step-by-step. The similarity between two samples is usually given by the Euclidean distance, and a “distance” can be represented by the “difference” between analytical values from both samples. The process of forming and joining clusters is repeated until a single cluster containing all samples is obtained, and the result can be displayed graphically in a dendrogram or tree diagram [24]. The dendrogram provides a visual summary of the clustering process, presenting a picture of the groups and its proximity with a dramatic reduction in dimensionality of the original data [13].

In this work, HCA was presented on the normalized data set using the Ward’s method as agglomeration technique and squared Euclidean distance as a measure of similarity. The Ward’s method employs an analysis of variance approach to evaluate the distances between clusters, attempting to minimize the sum of squares of any two clusters that can be formed at each step. The Euclidean distance (linkage distance) is reported as $D_{\text{link}}/D_{\text{max}}$, which represents the quotient between the linkage distance divided by the maximal distance. The quotient is usually multiplied by 100 as a way to standardize the linkage distance represented by the vertical-axis [8, 12, 17].

2.5 Principal component analysis/factor analysis

Principal component analysis (PCA) provides information on the most meaningful parameters, which describe the whole data set rendering data reduction with minimum loss of original information [7, 19, 23]. PCA starts with the covariance matrix describing the dispersion of the original variables (measured parameters), and extracting the eigenvalues and eigenvectors. An eigenvector is a list of coefficients (loadings or weightings) by which we multiply the original correlated variables to obtain new uncorrelated (orthogonal) variables, called principal components (PCs), which are weighted linear combinations of the original variables [7, 23]. An eigenvalue gives a measure of the significance of the PC; thus, the PCs with the highest eigenvalues are the most significant. Eigenvalues of 1.0 or greater are considered significant [17].

Factor analysis (FA) follows PCA. The main purpose of FA is to reduce the contribution of less significant variables in order to simplify even more of the data structure coming from PCA. This last purpose can be achieved by rotating the axis defined by PCA, according to well-established rules, and constructing new groups of variables, also called varifactors (VFs). It should be noted that a PC is a linear combination of observable water quality variables, while a VF can include unobservable, hypothetical, “latent” variables [7, 23]. The factor loadings express the correlation between the original variables and the newly formed varifactors. The VF loadings can be used to determine the relative importance of a variable as compared to other variables in a factor and do not reflect the importance of the factor itself [3]. Classification of factor loadings is “strong”, “moderate”, and “weak”, corresponding to absolute loading values of >0.75, 0.75−0.50 and 0.50−0.30, respectively [25].

3 Results and discussion

Table 2 summaries briefly the maximum, minimum and mean values, standard deviation (Std. Dev.) and coefficient variation (CV) of the 12 measured parameters in the river water samples from the 10 sampling sites in the Nenjiang River basin. It must be noticed that high dispersion of most variables (high standard deviation and coefficient variation), which indicates variability in chemical composition between samples, thus pointing to the presence of spatial and temporal variations caused likely by polluting sources and/or climatic factors. Recommended guide levels of these variables allowed by the Environmental Quality Standards for Surface Water (EQSSW, GB3838 — 2002, National Environmental Protection Agency of China 2002) are included in Table 2. It must be emphasized that average concentrations of some variables are within the acceptable limit of the Grade III standard, therefore this water resource is adequate for human consumption or industrial purposes.

3.1 Temporal similarity and period grouping

Temporal HCA generated a dendrogram (Fig. 3), grouping 12 months into three clusters with significant differences at $(D_{\text{link}}/D_{\text{max}})\times 100<64$.

Cluster 1 (the first period) includes January, February, March, November and December, which corresponds to the low flow (LF) period. In this period, all streams are so severely icebound that the depth of the ice layer above the water body usually reaches 0.8−1.5 m. Cluster 2 (the second period) includes April, May and June, which approximately corresponds to the typical mean flow (MF) period in Northeast China. It is also the

Table 2 Statistical descriptive of water quality in Nenjiang River basin, China

Parameter	Min	Max	Mean	Std. Dev.	CV/%	Environmental Quality Standards for Surface Water (GB3838—2002)				
						I	II	III	Ⅳ	V
Temp	4.98	26.05	16.52	7.24	43.80	–	–	–	–	–
SD	0.05	1.00	0.32	0.21	65.56	–	–	–	–	–
pH	7.48	10.32	8.19	0.42	5.15	6.5−8.5	6.5−8.5	6.5−8.5	6.5−8.5	6.5−8.5
DO	5.24	12.28	8.62	2.09	24.23	>7.5	6	5	3	2
EC	0.08	20.80	5.43	6.69	123.31	–	–	–	–	–
NH_3-N	0.32	0.86	0.57	0.13	23.89	<0.15	0.5	1.0	1.5	2.0
TOC	2.40	7.19	5.72	1.07	18.64	–	–	–	–	–
COD	2.01	9.22	5.25	2.48	47.25	<2	4	6	10	15
BOD_5	0.00	5.62	2.52	2.16	85.62	<3	3	4	6	10
TN	0.57	1.17	0.91	0.18	20.26	<0.2	0.5	1.0	1.5	2.0
TP	0.05	0.18	0.11	0.04	32.87	<0.02	0.1	0.2	0.3	0.4
Chla	0.08	32.50	7.45	6.81	91.43	–	–	–	–	–

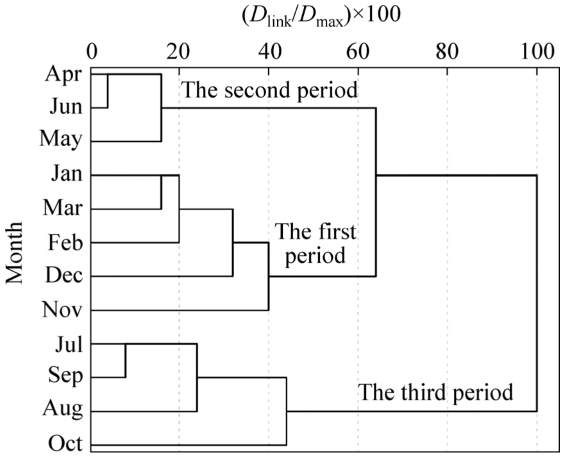

Fig. 3 Dendogram showing clustering of monitoring periods based on hierarchical clustering (Ward's method)

interim between non-icebound and icebound periods. In the upper basin the spring stream flow from snow melting takes up more than 10% of total stream flow and the percentage of that is smaller in the lower basin. Cluster 3 (the third period) includes July, August, September and October, which closely corresponds to the high flow (HF) period. The phenomena can be explained by the local summer flood.

Therefore, 12 months were divided into three different clusters by their hydrological characteristics (low, mean and high flows) rather than the traditional four seasons (spring, summer, autumn and winter). This temporal pattern of water quality is actually more reasonable because the winter in the Nenjiang River basin is so long that it could last nearly half a year; even the well-trained hydrologist may not be able to easily distinguish the months of each period by only reviewing the discharge record [26].

3.2 Temporal variations of river water quality

Temporal variations of the water quality parameters were first evaluated through a hydrological period parameter correlation matrix, using the Spearman non-parametric correlation coefficient (*R*). To do this, a specific integer number was assigned to each period (LF, 1; MF, 2; HF, 3). Then, Spearman correlation was established between all of the water quality parameters and the ordinal variables [6, 13, 17].

The results show that the water temperature exhibits the highest correlation coefficient (R=0.764) with the period and a very significant p-level (0.000). In addition to the temperature, we observed six additional parameters having significant correlation with the period (p<0.05): DO (R=−0.680), SD (R=−0.475), COD (R=0.387), NH_3-N (R=0.342), Chla (R=−0.332), and BOD_5 (R=0.294). So far, these parameters can be taken as representing the major source of temporal variations in water quality.

These correlations in various water quality parameters can be explained in terms of the climatologic and hydrologic characteristics associated with the period. It is evident that the water temperature reflects the atmospheric temperature, and that this parameter presents the most significant difference between the three periods. From the temperature difference, changes are expected in the DO. The negatively correlation between the period and SD can also be explained in terms of the increased quantity of eroded material and urban runoff expected while raining (HF period). Flow rate is negatively correlated to most variables, since an increase in flow rate caused dilution of contaminants. The correlation observed with COD and BOD_5 could be a result of increased anthropogenic activities and river flow during the HF period. In the LF period (such as winter), there are not many agricultural activities going on in this area, which leads to low load of contaminants; in the HF period (such as summer), there is a general increase in agricultural activities, producing a greater contribution of organic matter, through increased runoff and severe erosion (Fig. 4). However, the monitoring results show that there is a high concentration of NH_3-N during the LF period, while there is a low concentration during the HF period in the study area (Fig. 4). Temporal variation of NH_3-N is due to seasonal change of river hydrology because the degradation of NH_3-N is highly correlated with temperature. Thus, seasonal variation in the concentration of NH_3-N can be attributed to natural seasonal influences [26]. Non-significant correlation of other parameters with period indicates the contribution of anthropogenic sources in the catchment areas.

Temporal PCA/FA was further applied to the standardized data sets containing 12 variables, separately for three periods, viz., LF, MF and HF period, as delineated by temporal HCA techniques, to compare the compositional pattern between water quality parameters and to identify the factors influencing them. The KMO result for the LF, MF and HF period were 0.688、0.778 and 0.869, and Bartlett's sphericity test was significant (0.000, p<0.05), showing that PCA/FA could be considered appropriate and useful to provide significant reduction in data dimensionality.

PCA renderes four PCs for the LF period, three PCs for the MF period and two PCs for the HF period with eigenvalues > 1, explaining 85.8%, 79.6% and 72.0% of the total variance in respective water quality data sets. Equal numbers of varifactors (VFs) are obtained through the FA performed on the PCs. Results of FA including factor loadings, eigenvalues and total and cumulative variance values are presented in Table 3.

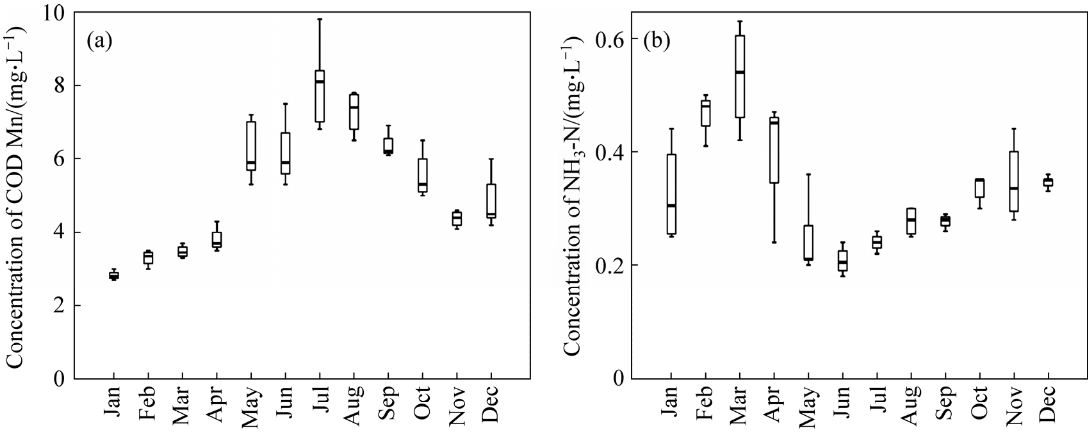

Fig. 4 Boxplot of COD Mn (a) and NH_3-N concentration (b) in Nenjiang River basin, China (2012−2013)

Table 3 Factor loadings matrix and explained variance of water quality parameters in three periods

Parameter	LF period				MF period			HF period	
	VF1	VF2	VF3	VF4	VF1	VF2	VF3	VF1	VF2
Temp	**0.935**	0.191	−0.108	−0.198	*−0.699*	*−0.639*	−0.113	**−0.791**	−0.215
SD	**−0.806**	0.456	−0.075	−0.101	0.102	**0.929**	−0.075	0.480	0.340
pH	0.004	−0.101	**0.819**	−0.289	−0.425	*0.691*	−0.128	*−0.739*	−0.167
DO	**−0.819**	−0.306	−0.108	0.000	0.497	*0.669*	−0.153	*0.680*	0.392
EC	0.232	−0.204	0.478	*−0.584*	**−0.871**	−0.211	0.056	**0.962**	0.027
NH_3−N	0.182	**0.917**	0.034	0.149	−0.436	−0.320	*0.723*	0.376	**0.951**
TOC	−0.001	0.064	0.015	**0.916**	*0.576*	0.420	0.320	0.378	**0.772**
COD	**0.915**	0.323	−0.033	0.022	**0.965**	0.160	−0.105	**0.863**	0.329
BOD_5	**0.932**	0.255	−0.160	−0.059	**0.912**	0.096	−0.242	0.110	**−0.888**
TN	0.283	**0.912**	0.065	0.062	**0.833**	0.056	0.312	0.265	**0.857**
TP	−0.076	0.288	**0.807**	0.213	0.162	−0.005	**0.862**	**0.913**	0.276
Chla	**−0.935**	−0.001	−0.137	0.085	**0.823**	−0.184	0.018	**−0.905**	−0.068
Eigenvalue	4.948	2.318	1.630	1.397	5.368	2.592	1.598	5.289	3.354
Total variance/%	41.236	19.318	13.581	11.640	44.737	21.597	13.313	44.072	27.953
Cumulated/%	41.239	60.553	74.134	85.774	44.737	66.334	79.647	44.072	72.024

Note: Bold values are coefficients higher than or equal to 0.75; italic values are higher than or equal to 0.5 (with significance level of 0.05).

For the dataset pertaining to the LF period, among four VFs, VF1 (41.236% of total variance) has strong positive loadings on BOD_5, COD and Temp, and strong negative loadings on Chla, DO and SD. In the frozen period, a thick surface ice layer would prevent nonpoint sources pollution from agricultural activities like animal breeding, agricultural fertilizers, and soil erosions. Thus, this organic factor mainly represents the contribution of point sources, such as municipal and industrial discharge pipes and wastewater treatment plants. Furthermore, DO decreases in the LF period, probably related to 1) high levels of dissolved organic matter consuming large amounts of oxygen, and 2) the thick surface ice layer would preventing photosynthesis and water re-aeration from the atmosphere when the river is icebound in the frozen period. The high loading of temperature is associated with seasonal variation, thus showing that only climate and seasonality are responsible for variations in water temperature. VF2 (19.318% of total variance) has strong positive loadings on NH_3-N and TN. VF3 (13.581% of total variance) has strong positive loadings on pH and TP. VF2 and VF3 reflect the nutrient pollution of nitrogen and phosphorus. The excess nutrient in the LF period was mainly from point sources, such as town sewage. VF4 (11.640% of total variance) has strong positive loadings on TOC and moderate negative loadings on EC; these parameters are indicators of organic pollution from industries activities.

For the data set pertaining to the MF period, among three VFs, VF1 (44.737% of total variance) has strong positive loadings on COD, BOD_5, TN, and Chla and moderate positive loadings on TOC, whereas strong negative loadings on EC and moderate negative loadings on Temp. This factor can be explained as oxygen-consuming organic and inorganic nutrients pollution. This factor may be related to anthropogenic pollution of industrial, domestic and agricultural source. VF2 (21.597% of total variance) has strong positive loadings on SD and moderate positive loadings on pH and DO, whereas moderate negative loadings on Temp. Strong loading on these parameters could have been due to anthropogenic activities through road construction, clearing of lands, and runoff taking place near the study area. Meanwhile, this factor also indicates a strong effect of soil erosion from agricultural fields especially when the river flow increases due to snow melting at the river sources. The negative loadings of SD and Temp can be explained taking into account that particles suspended in water may absorb heat in the sunlight, hence raising water temperature. VF3 (13.313% of total variance) has strong positive loadings on TP and moderate positive loadings on NH_3-N. This nutrient factor represents the non-point pollution of the river. Non-point sources of nitrogen in this region mainly contain agricultural runoff, as well as natural decomposition organic and geologic deposits. TP also ascribes to the runoff from phosphorous fertilizers and soil erosion.

For the data set pertaining to the HF period, among two VFs, VF1 (44.072% of total variance) has strong positive loadings on EC, COD, and TP, and moderate positive loadings on DO, whereas strong negative loadings on Chla and Temp, and moderate negative loadings on pH. This factor can be explained as high levels of organic matter and inorganic nutrients. High loadings on these parameters indicate that the river is heavily polluted due to anthropogenic activities through point and non-point sources. Point sources such as wastewater from domestic sewage, wastewater treatment plants and industries effluences. Impact of non-point sources of pollution especially agricultural activities like animal breeding, agricultural fertilizers, and soil erosions are dominant. VF2 (22.953% of total variance) has strong positive loadings on TN, NH_3-N and TOC, and strong negative loadings on BOD_5. High concentration of nitrogen and organic matter in river can originate from a number of sources, such as domestic sewage and agriculture cultivation. A large amount of nitrogenous fertilizers is applied on these farmlands, which is finally released into the river ecosystem and deteriorates the river water quality. In the HF period, flushing of overland runoff due to floods is the main and nearly constant sources of organic matter and nutrient in river water, so that dilution of pollutants caused by an increasing in river flow is less evident.

3.3 Spatial similarity and site grouping

Spatial HCA rendered a dendrogram, grouping 10 sampling sites into three distinct clusters with significant differences at $(D_{\text{link}}/D_{\text{max}})\times 100 < 56$ (Fig. 5).

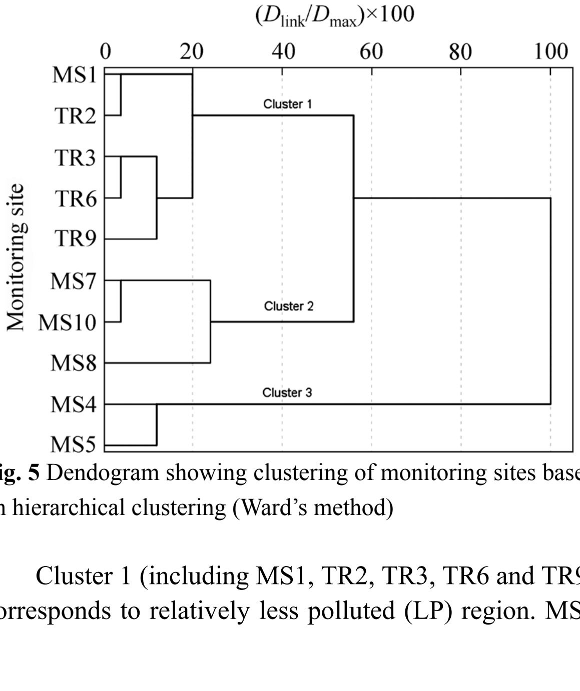

Fig. 5 Dendogram showing clustering of monitoring sites based on hierarchical clustering (Ward's method)

Cluster 1 (including MS1, TR2, TR3, TR6 and TR9) corresponds to relatively less polluted (LP) region. MS1 is located in the headwater of the Nenjiang River in an area that has extensive forest cover and limited human activities, thus the water quality is optimal. TR2 and TR3 are both located in the upstream and midstream of the river in areas that experience no industrial activity and have relatively small human population; thus water quality is found to be only less polluted by agricultural practices. Relatively low concentration of all monitored water parameters were observed in TR6, possibly attributed to high water flow and long stream length in the Yalu River. TR9, located in the Hudaer River, has a medium-size reservoir, which the water quality was in a better condition.

Cluster 2 (containing MS7, MS8 and MS10) corresponds to moderately polluted (MP) region. In cluster 2, these sites are situated at the most downstream area of the Nenjiang River basin. Although the direct discharged domestic wastewater and industrial effluents and surface runoff from villages contaminated the river, the water quality corresponds to moderately polluted, which indicates the existence of the self-purification and assimilative capacity of the river as it flows downstream.

Cluster 3 (comprising MS4 and MS5) corresponds to highly polluted (HP) region. In cluster 3, these sites are located at the Qiqihar City region. Qiqihar City is nearly the most developed region in the Nenjiang River basin, with a population of about 5.71 million (Heilongjiang Statistical Yearbook 2010). There are many chemical, pharmaceutical, petroleum chemical, iron-steel and electroplating factories. These sites received large amounts of pollution from various point sources, such as wastewater treatment plants, domestic wastewater and industrial effluents, and showed the highest average concentrations of all monitored water parameters. Moreover, a large number of livestock breeding farms without wastewater treatment facilities also located near the river.

3.4 Spatial variations in river water quality

Spatial PCA/FA is performed on the correlation matrix of rearranged normalized datasets separately for the three regions (LP, MP and HP) to compare the compositional pattern between analyzed water samples and identify the source influencing each one. In this study, the KMO result for the HP, MP and LP regions were 0.729, 0.671 and 0.631, and Bartlett's Sphericity test was significant (0.00, p<0.05), indicating that PCA/FA could be considered appropriate and useful to provide significant reduction in data dimensionality.

PCA yield three PCs for the LP and HP regions and four PCs for the MP region with eigenvalues>1, explaining 79%, 88% and 87% of the total variance in respective water quality data sets. Equal numbers of VFs are obtained through FA performed on the PCs. Results of FA including factor loadings, eigenvalues and total and cumulative variance values are presented in Table 4.

For the dataset pertaining to water quality in the LP sites, among three VFs, VF1 (30.305% of total variance) has strong positive loadings on DO and SD, and moderate positive loadings on TOC and TP, whereas strong negative loadings on Temp and moderate negative loadings on pH. This factor explains the erosion from

Table 4 Factor loadings matrix and explained variance of water quality parameters in three regions

Parameter	LP region			MP region				HP region		
	VF1	VF2	VF3	VF1	VF2	VF3	VF4	VF1	VF2	VF3
Temp	**−0.941**	0.197	−0.037	**−0.938**	0.213	−0.082	0.187	0.078	**−0.957**	−0.056
SD	**0.800**	0.098	0.101	**0.908**	−0.287	−0.126	−0.126	**−0.884**	0.315	0.105
pH	*−0.641*	0.079	−0.282	0.049	**−0.925**	0.147	−0.096	*−0.711*	*−0.568*	0.150
DO	**0.932**	0.103	−0.084	**0.950**	0.118	−0.116	0.021	−0.141	**0.906**	−0.314
EC	0.021	−0.149	**0.952**	0.398	−0.343	−0.133	**−0.779**	−0.453	0.490	*−0.712*
NH_3-H	−0.037	−0.421	*0.650*	−0.286	0.023	**0.794**	0.174	0.237	*−0.549*	*−0.667*
TOC	*0.649*	*0.540*	0.225	0.264	0.271	0.378	*0.544*	*0.602*	*0.538*	0.132
COD	−0.028	**0.993**	0.027	−0.112	**0.775**	*0.531*	0.196	**0.922**	0.130	0.302
BOD_5	−0.225	**0.927**	−0.270	−0.235	**0.774**	0.450	0.219	**0.816**	−0.073	*0.503*
TN	0.220	0.090	**0.810**	−0.111	0.441	*0.700*	0.364	**0.911**	−0.008	0.290
TP	*0.539*	0.205	*0.688*	*0.536*	0.022	**0.759**	−0.161	0.227	**0.881**	0.116
Chla	−0.137	−0.126	**−0.792**	−0.124	0.027	0.042	**0.959**	0.359	−0.027	**0.881**
Eigenvalue	3.637	3.210	2.659	3.302	2.585	2.401	2.157	4.463	3.786	2.322
Total variance/%	30.305	26.748	22.157	27.514	21.544	20.007	17.975	37.191	31.554	19.349
Cumulate/%	30.305	57.053	79.210	27.514	49.058	69.065	87.040	37.191	68.745	88.094

Note: Bold values are coefficients higher than or equal to 0.75; italic values are higher than or equal to 0.5 (with significance level of 0.05).

upland areas during rainfall events and the positive correlation with TOC and TP indicates the loading of partially decayed organic matters from forested areas. This assumption is reasonable, as the water quality in this region is good and land use activities are mostly limited to agriculture and forest areas. The inverse relationship between Temp and DO is also a natural process because warmer water becomes saturated more easily with oxygen and it can hold less dissolved oxygen. VF2 (26.748% of total variance) has strong positive loadings on COD and BOD_5, and moderate positive loadings on TOC. BOD_5, COD and TOC are considered organic factors and may be interpreted as representing influences from nonpoint sources such as agricultural activities and forest areas. VF3 (22.157% of total variance) has strong positive loadings on EC and TN, and moderate positive loadings on NH_3-N and TP, whereas strong negative loadings on Chla. The presence of nitrogen and phosphorus are related to the influence of domestic waste and agricultural runoff.

For the dataset pertaining to water quality in the MP sites, among four VFs, VF1 (27.514% of total variance) has strong positive loadings on DO and SD, and moderate positive loadings on TP, whereas strong negative loadings on Temp. Different from the LP region, this factor is correlated with suspended solids, which is related to the discharge from urban development areas involving clearing of lands, the erosion of road edges due to surface runoff, as well as agricultural runoff. The conversion of land use from forestal or agricultural to urban has indeed caused large negative impacts to the ecosystem of the Nenjiang River basin in the form of mud flood, land slide, and river floods. VF2 (21.544% of total variance) has strong positive loadings on COD and BOD_5 and strong negative loadings on pH. As explained above, this factor is related to organic pollution and suspected to come from point pollution sources such as sewage treatment plants and industrial effluents. Organic matters in river water consumes large amount of oxygen, and as the amount of available DO decreases, they undergo anaerobic fermentation processes leading to the production of ammonia and organic acids. Hydrolysis of these acidic materials causes a decrease of water pH values. VF3 (20.007% of total variance) has a high and positive load of NH_3-N and TP, and moderate positive loadings on TN and COD. This nutrient factor represents the point and non-point pollution of the river. Point sources include municipal waste treatment plants, industrial operations, and large, confined livestock operations. Nonpoint sources comprise soil erosion and water runoff from cropland, lawns and gardens, private waste treatment systems, urban areas, small livestock confinement operations, etc. VF4 (17.975% of total variance) has strong positive loadings on Chla and moderate positive loadings on TOC, whereas strong negative loadings on EC. This factor is suspected to originate from agricultural fields, where irrigated horticultural crops are grown and the use of inorganic fertilizers (usually as ammonium nitrate) is rather frequent. But this pollution may also originate from the decomposition of nitrogen containing organic compounds via degradation process of organic matters, such as proteins and urea occurring in municipal wastewater discharges.

Lastly, for the data set representing the HP sites, among total three VFs, VF1 (37.191% of total variance) has strong positive loadings on COD, TN and BOD_5, and moderate positive loadings on TOC, whereas strong negative loadings on SD and moderate negative loadings on pH. This factor indicates that the river is heavily polluted with organic pollutants. In the HP region, this pollution comes mostly from point sources such as discharge from wastewater treatment plants, domestic wastewater and industrial effluents. The presence of nutrient in this region, also possibly contributed by pollution loadings from livestock farms, attributed to the absence of a treatment system. The negative factor loading of SD on this factor can be attributed to run-off from fields with high load of soil and waste disposal activities. VF2 (31.554% of total variance) has strong positive loadings on TP and DO and strong negative loading on Temp. This factor can be explained as the phosphorous pollutions of the stream. Point sources of the phosphorus such as wastewater from the phosphorous chemical industry and non-point sources like animal breeding, agricultural fertilizers, and soil erosions constitute the pollution commonly found in this region. VF3 (19.349% of total variance) has strong positive loadings on Chla and moderate positive loadings on BOD_5, whereas moderate negative loadings on EC and NH_3-N. The presence of nitrogen is due to agricultural runoff such as livestock waste and fertilizers, industrial effluents, municipal sewage, and existing sewage treatment plants because nitrogen is an important component of detergents.

4 Conclusions

1) In this case work, different multivariate statistical techniques are used to evaluate temporal and spatial variations in surface water quality of the Nenjiang River basin. Hierarchical cluster analysis (HCA) renders good results to evaluate both temporal and spatial differences. HCA groups 12 months into three periods (LF, MF and HF) and classifies 10 monitoring sites into three regions (LP, MP and HP) based on the similarity of water quality characteristics. It offers reliable classification of the surface water in the Nenjiang River basin that could help

to design an optimal future monitoring strategy. According to spatial HCA, the number of monitoring sites may be decreased and only chosen from clusters 1, 2 and 3. Similarly, according to temporal HCA, the monitoring frequency may decrease and the monitoring period could be selected depending on their hydrological characters (LF, MF and HF), rather than the traditional seasons. Thus, the HCA can facilitate comparisons among different localities or periods and can be a useful tool for optimizating the water quality monitoring strategy.

2) According to the results from temporal PCA/FA, we can observe that beside geochemical aspects, seasonal regime of the Nenjiang River basin water quality is controlled by three important hydrologic/ anthropogenic processes: ① point sources pollution, such as wastewater from domestic sewage and wastewater treatment plants in LF, MF and HF periods; ② impact of non-point sources of pollution especially agricultural activities, that are dominant during the HF and MF periods; ③ flushing of overland runoff, due to floods occurring in HF period. In general, temporal effects were associated to variations of river flow rate which cause dilution of pollutants and hence variations in water quality. Having the mentioned processes functioning over the watershed, relevant management policies and actions need to implement. Of the best ways to reduce the concentrations of chemical compositions of river water, establishment of riparian vegetation is recommended.

3) The result of spatial PCA/FA indicates that the parameters responsible for water quality spatial variation were mainly related to suspended solids (natural sources: soil erosion), organic pollution and nutrients (non-point sources: animal husbandry and agriculture activities) in relatively LP region; suspended solids (non-point sources: clearing of lands, surface runoff, agricultural runoff), oxygen-consuming organic pollution (point sources: industries and domestic wastewater), nutrients (non-point sources: agriculture activities, organic decomposition and geologic deposits) in MP region; and oxygen-consuming organic pollution (point source: domestic sewage, wastewater treatment plants and industrial effluences), nutrients (non-point sources: agricultural activities, runoff in soils) in HP region. Consequently, the PCA/FA could be useful for evaluation of potential environmental hazards in each region. Therefore, different measures can be carried out to control the water pollution sources in different regions. The HP region should pay more attention to these point sources of pollution that are also significant latent pollution sources in MP region. Non-point sources of pollution are a serious environmental problem throughout the basin. Agricultural activities like cultivation or aquatic breeding without advanced techniques bring tremendous nutrients and organic pollution. Thus, the priority is to develop advanced techniques for decreasing non-point sources of pollution in these three regions.

4) The application of HCA and PCA/FA has achieved meaningful classification based on temporal and spatial criteria. This work reinforces the fact that multivariate statistical methods including HCA and PCA/FA can be applied to interpret complex datasets of water quality, understand temporal and spatial variations in water quality, and identify latent pollution sources/ factors. Therefore, this evaluation work can help managers identify the main sources of pollution in different regions and in different periods so as to determine their priorities for improving water quality. Since multivariate statistical methods are easily applied to water quality data, using them can be a practical approach to environmental impact assessment.

References

[1] JAEHNIG S C, CAI Q. River water quality assessment in selected Yangtze tributaries: Background and method development [J]. Journal of Earth Science, 2010, 21(6): 876−881.

[2] WANG X Z, CAI Q H, YE L, QU X D. Evaluation of spatial and temporal variation in stream water quality by multivariate statistical techniques: A case study of the Xiangxi River basin, China [J]. Quaternary International, 2012, 282: 137−144.

[3] OUYANG Y, NKEDI-KIZZA P, WU Q T, SHINDE D, HUANG C H. Assessment of seasonal variations in surface water quality [J]. Water Research, 2006, 40: 3800−3810.

[4] MUSTAPHA A, ARIS A Z, JUAHIR H, RAMLI M F, KURA N U. River water quality assessment using environmentric techniques: case study of Jakara River Basin [J]. Environmental Science and Pollution Research, 2013, 20: 5630−5644.

[5] OMERNIK J M. Ecoregions of the conterminous United States [J]. Annals of the Association of American Geographers, 1987, 77: 118−125.

[6] ZARE G A, SHEIKH V, SADODDIN A. Assessment of seasonal variations of chemical characteristics in surface water using multivariate statistical methods [J]. International Journal of Environmental Science and Technology, 2011, 8(3): 581−592.

[7] VEGEA M, PARDO R, BARRADO E, DEBAN L. Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis [J]. Water Research, 1998, 32: 3581−3592.

[8] 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 [J]. Analytica Chimica Acta, 2005, 35: 3581−3592.

[9] KAZI T G, ARAIN M B, JAMALI M K, JALBANI N, AFRIDI H I, SARFRAZ R A, BAIG J A, SHAH A Q. Assessment of water quality of polluted lake using multivariate statistical techniques: A case study [J]. Ecotoxicology and Environmental Safety, 2009, 72: 301−309.

[10] DIXON W, CHISWELL B. Review of aquatic monitoring program design [J]. Water Research, 1996, 30: 1935−1948.

[11] KOKLU R, SENGORUR B, TOPAL B. Water quality assessment using multivariate statistical methods—A case study: Melen River system (Turkey) [J]. Water Resource Management, 2010, 24: 959−978.

[12] SINGH K P, MALIK A, MOHAN D, SINHA S. Multivariate

statistical techniques for the evaluation of spatial and temporal variations in water qualityof Gomti River (India)—A case study [J]. Water Research, 2004, 38: 3980−3992.

[13] ALBERTO W D, PILAR D M D, VALERIA A M, FABIANA P S, CECILIA H A, ANGELES B M D L. Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. A case study: Squia River Basin (Cordoba-Argentina) [J]. Water Research, 2000, 35: 2881−2894.

[14] MCNEIL V H, COX M E, PREDA M. Assessment of chemical water types and their spatial variation using multi-stage cluster analysis, Queensland, Australia [J]. Journal of Hydrology, 2005, 310(1/2/3/4): 181−200.

[15] PANDA U C, SUNDARAY S K, RATH P, NAYAK B B, BHATTA D. Application of factor and cluster analysis for characterization of river and estuarine water systems—A case study: Mahanadi River (India) [J]. Journal of Hydrology, 2006, 331(3/4): 434−445.

[16] KANNEL P R, LEE S, KANEL S R, KHAN S P. Chemometric application in classification and assessment of monitoring locations of an urban river system [J]. Analytica Chimica Acta, 2007, 582: 390−399.

[17] SHRESTHA S, KAZAMA F. Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji River basin, Japan [J]. Environmental Modelling & Software, 2007, 22: 464−475.

[18] NAJAR I A, KHAN A B. Assessment of water quality and identification of pollution sources of three lakes in Kashmir, India, using multivariate analysis [J]. Environmental Earth Sciences, 2012, 66: 2367−2378.

[19] BIERMAN P, LEWIS M, OSTENDORF M, TANNER J. A review of methods for analyzing spatial and temporal patterns in coastal water quality [J]. Ecological Indicators, 2011, 11: 103−114.

[20] FENG X Q, ZHANG G X, YIN X R. Hydrological responses to climate change in Nenjiang River Basin, Northeastern China [J]. Water Resources Management, 2011, 25: 677−689.

[21] ZHANG B, SONG X F, ZHANG Y H, HAN D M, TANG C Y, YU Y L, MA Y. Hydrochemical characteristics and water quality assessment of surface water and groundwater in Songnen plain, Northest China [J]. Water Research, 2012, 46: 2737−2748.

[22] JOHNSON R A, WICHERN D W. Applied multivariate statistical analysis [M]. New Jersey: Prentice-Hall Int, 1992.

[23] HELENA B, PARDO R, VEGA M, BARRADO E, FERNANDEZ J M, FERNANDEZ L. Temporal evaluation of groundwater composition in an alluvial aquifer (Pisuerga river, Spain) by principal component analysis [J]. Water Research, 2000, 34: 807−816.

[24] MCKENNA J E Jr. An enhanced cluster analysis program with bootstrap significance testing for ecological community analysis [J]. Environmental Modelling & Software, 2003, 18(2): 205−220.

[25] LIU C W, LIN K H, KUO Y M. Application of factor analysis in the assessment of groundwater quality in a Blackfoot disease area in Taiwan [J]. The Science of the Total Environment, 2003, 313: 77−89.

[26] WANG Y, WANG P, BAI Y J, TIAN Z X, LI J W, SHAO X, MUSTAVICH L F, LI B L. Assessment of surface water quality via multivariate statistical techniques: A case study of the Songhua River Harbin region, China [J]. Journal of Hydro-environment Research, 2013, 7: 30−40.

(Edited by DENG Lü-xiang)