Desalination and Water Treatment
www.deswater.com

doi: 10.5004/dwt.2017.20970

96 (2017) 255–266
November

Spatiotemporal features and delineation of water quality control zones for Taipei Water Resources District with multivariate manners

Yu-Jie Chang[a], I-Cheng Chang[b], Tai-Yi Yu[c,*]

[a]*Department of Earth and Life Science, University of Taipei, No.1, Ai-Guo West Road, Taipei 10048, Taiwan, email: yjchang@utaipei.edu.tw*
[b]*Department of Environmental Engineering, National Ilan University, No.1, Sec. 1, Shennong Road, Yilan City, Yilan 26047, Taiwan, email: icchang@niu.edu.tw*
[c]*Department of Risk Management and Insurance, Ming Chuan University, Taipei 11103, Taiwan, email: yti@mail.mcu.edu.tw*

Received 27 January 2017; Accepted 26 May 2017

ABSTRACT

This research adopts two methods of multivariate statistical analysis (MSA), principal component analysis (PCA) and cluster analysis (CA), to analyze the water quality parameters (WQPs) monitoring results for evaluating the dominant factors on the rivers water quality and the areas which should be protected carefully. The combination of PCA and CA provides a better technique to classify the water quality control zones. Although PCA is an effective tool to categorize the monitoring stations, it cannot conduct complex dimensional classification on all of the monitoring stations and parameters; whereas, CA can help to determine the correlations between different monitoring stations via the WQPs monitoring results and then provides a more reasonable classification numbers for further watershed management. In this research, 23 monitoring stations were classified into four water quality control zones by using PCA and CA methods. The results from PCA in various water quality control zones indicate that the amounts of total coliform (TC) can lead to various correlation with various WQPs based on the characteristics of regions and pollutant sources. By applying CA to further classify the WQPs of the monitoring station for midstream of Nanshi River, analysis of variance (ANOVA) tests found only the mean values of monitoring WQPs indices for TC and dissolved oxygen (DO) have significant differences. In terms of the water quality in this area, the wastewater from hot springs usages might cause 17% of the midstream of Nanshi River monitoring stations (Cluster A) to rise their TC values and slightly decrease both DO and pH values. In this region, TC is the WQPs indicator with the highest impact resulted from hot spring wastewater. Additionally, by applying PCA and CA, the correlation of WQPs and the effects that hot spring wastewater have on water quality can be further investigated.

Keywords: Principal component analysis; Cluster analysis; River water quality parameters

1. Introduction

The variables which affect river basin water quality are complex and diverse. Therefore, reducing the number of variables and finding the key variables are not only cost-effective, but also it can provide understanding of the characteristics of water pollution patterns and even effective strategies for water quality management. Water quality of river basin is the integrated result of sources, hydraulic factors, geochemistry and other complicated environmental variables. Due to its complexity and diversification, it is not easy to use one single variable to analyze the correlation or characteristics of water quality parameters (WQPs). Multivariate statistical analysis (MSA) is a quantitative analysis method suitable for analyzing more than two variables. It is often used to simplify variables, to establish cause–effect relationships and to

* Corresponding author.

Presented at the 9th International Conference on Challenges in Environmental Science & Engineering (CESE-2016), 6–10 November 2016, Kaohsiung, Taiwan.

understand the correlation between each other. It is a suitable statistical tool for simplifying variables which affect the water quality of river basins effectively.

The application of principal component analysis (PCA) in the field of environmental management or environmental pollution is comprehensive, particularly environmental problems related to air and water. The PCA manner was widely employed to simplify the number of complicated variables [1–3], to analyze the correlation between pollutant concentrations and physical parameters [4,5], to identify the cause–effect relationships between pollutants and sources [6,7] and to assess the validity and concentration trends of monitoring data [8,9]. Eder et al. [10] stressed three main advantages of applying PCA for spatial delineation: (1) providing spatial delineation with statistical and physical significance; (2) understanding the distribution and characteristics of pollutant concentrations in the sub-regions; and (3) providing the common characteristics of most monitoring stations, overcoming the bottleneck of being able to interpret single monitoring station. In terms of air pollution, Eder [11] and Eder et al. [12] adopted PCA to analyze the concentration of sulfuric acid (40 stations) and hourly ozone values (77 stations), and results classified seven and six sub-regions with the interpreted concentration variation of up to 74% and 64%, respectively. Vardoulakis and Pavlos [13] applied PCA and regression analysis to quantify the contribution of non-combustion sources to background concentrations of PM10; whereas, in terms of water pollution, Koklu et al. [14] employed PCA technique to evaluate high–low flow periods correlations of WQPs and extract dominant factors in assessing variations of river water quality. Olsen et al. [15] found that PCA was the most frequently used MSA method for assessing variations of water quality in river basins. Researchers often use PCA to identify and describe spatial patterns of water quality, using geochemical processes, hydraulic programs and locations of sources of pollution to interpret the spatial features of water quality. Olsen et al. [15] also reviewed the application of PCA to assess variations of water quality in river basins and the results indicated that the application of PCA had no consistent procedures, but dependent on the variables of sampling design, data quality, types of WQPs, data pre-processing techniques, interpretation procedures and other related factors.

Regarding case studies that applied cluster analysis (CA) manner in the field of water pollution, Zhou et al. [16] analyzed 14 WQPs (2000–2004) from 27 monitoring stations on the East Coast of Hong Kong, and the WQPs data were classified into two clusters based on the degree of pollution (June–September as one, and the rest as the other) with CA method. Yang et al. [17] utilized various multivariate statistical methods including CA, discriminant analysis (DA), factor analysis (FA) and PCA were used to explain spatial and temporal patterns of surface water pollution in Lake Dianchi during the period of 2003–2007. Shrestha and Kazama [18] also applied four MSA methods to study the data of 12 WQPs from 13 monitoring stations in Fuij River Basin over the period of 1995–2002, and the results indicate that CA can help to categorize the data into three clusters according to the levels of pollution, namely, low-, medium- and high-level of pollution. By interpreting these three clusters via FA and PCA approaches, highest correlation of the polluted sources and the aforementioned three types of water quality pollution can be identified. Zhang et al. [19] used fuzzy membership analysis and MSA manners to classify and assess the water quality monitoring results of water quality for groundwater and surface water in Songnen Plain, China. The PCA and hierarchical cluster analysis (HCA) are both used to classify the different levels of water quality into four principal components and three clusters, respectively. Juahir et al. [20] utilized four MSA methods, hierarchical agglomerative cluster analysis (HACA), PCA, FA and DA, to investigate spatial variations of the most significant water quality variables and to determine the polluted sources and formed three spatial clusters with the HACA technique. Razmkhah et al. [21] applied PCA and CA to analyze 18 WQPs from 18 monitoring stations in order to study the effects of anthropogenic pollution on the water quality of Jajrood River in Iran, and found that PCA is suitable for explaining spatiotemporal variation of pollution concentration and CA is suitable for providing the classification and interpretation of clusters. Astel et al. [22] analyzed a large number of chemical WQPs by using PCA, CA and self-organizing maps (SOM) and found SOM clustering allows simultaneous observation of both spatial and temporal variations in river water quality. Three different patterns of monitoring sites were conditionally named as "tributary", "urban" and "background."

Taipei Water Resources Designated Area is the first protected area for water resources enacted by the Urban Planning Law in Taiwan. The catchment area is 717 km^2. The water quality management within this area will affect the safety and cleanness of water sources as well as water quality of upstream catchment area of Qingtan Weir, Xindian River, which directly affects the quality of drinking water supply system in Taipei area. Therefore, the identification of polluted sources, land use and water quality protection and management appear to be particularly important. The purposes of this research are to (1) identify possible polluted sources and analyze the variation of WQPs from water quality monitoring stations by applying PCA and CA manners, (2) understand the potential extent of influence by specific polluted sources on water quality with PCA and CA manners, (3) effectively and objectively demarcate water quality control zones and execute strategies for water quality management and (4) interpret the variation of WQPs that led to serious pollution by applying CA manner and analysis of variance (ANOVA) tests.

2. Materials and methods

The scope of Taipei Water Resources Designated Area (Fig. 1) encompasses three river streams: the tributaries of Xindian River (monitoring stations C1–C5), Nanshi River (monitoring stations B1–B6) and Beishi River (monitoring stations A1–A12). In total, there are 23 stations that routinely monitor the water quality of the rivers, with the following 12 WQPs being monitored monthly: water temperature (°C), pH, dissolved oxygen (DO), biochemical oxygen demand (BOD), suspended solids (SS), coliform (CFU/100 mL), ammonia nitrogen (NH_3–N), chemical oxygen demand (COD), conductivity (μS/cm), total phosphorus (TP), turbidity (NTU) and total organic carbon. The data taken for this analysis is the monthly monitoring data from 1987 to 2012, and the last four WQPs have only been monitored since 2008.

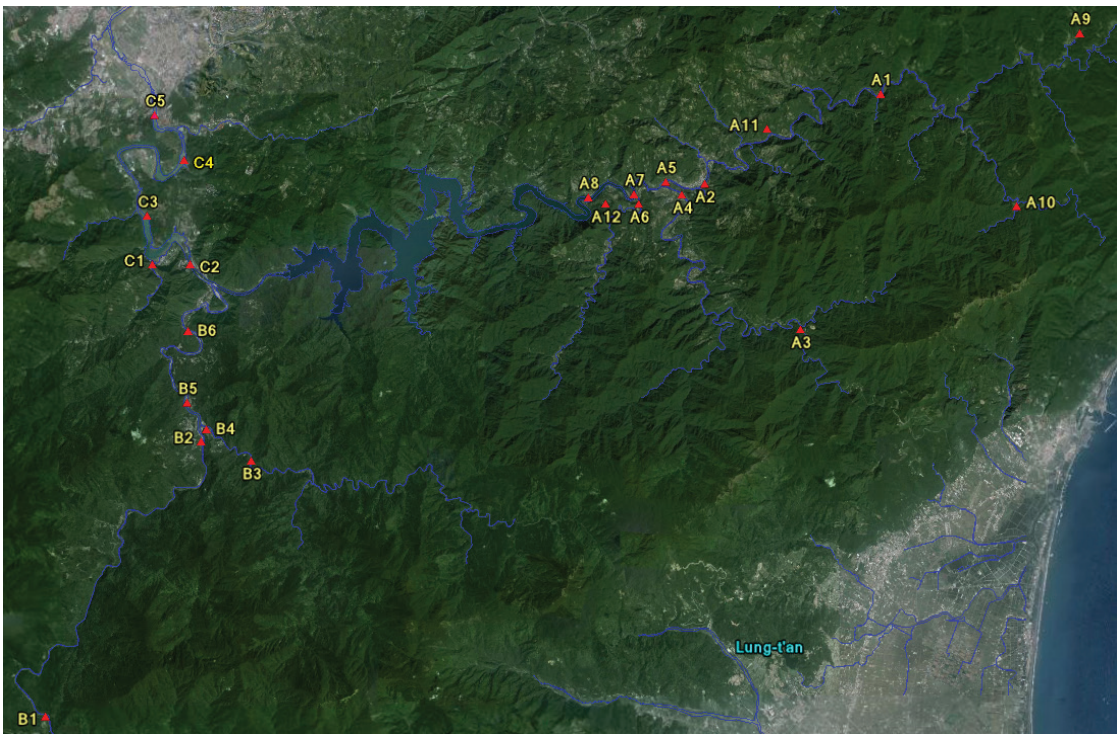

Fig. 1. Locations of monitoring stations. Monitoring stations over Xindian River (C1–C5); Nanshi River (B1–B6) and Beishi River (A1–A12).

Water quality index (WQI) use seven WQPs from WQPs as references, including pH, DO, BOD, SS, coliform (Most probable number (MPN)/100 mL), NH_3–N and TP. The weight of each parameter is based on Delphi experts' weighing [23]. According to the historical results of water quality, the first four dominant water quality indicators with high violation rates are: TP, the coliform (CFU/100 mL), SS and BOD. This research first performed the PCA and CA manners to analyze spatiotemporal patters of potential sources for the four dominant water quality indicators, and proposed the spatial delineation of water quality control zones. Second, the locations and pollution patterns of potential sources were identified with the results of PCA and CA approaches. Third, this research conducted the PCA manner to analyze the variation and features of concentration levels of water quality indicators in different water quality control zones. Finally, this research selected the water quality control zone with the most serious total coliform (TC) values and analyzed the variations and high levels of water quality indicators with CA approach and ANOVA tests.

2.1. Principal component analysis

Researches [24,25] have indicated that unrotated PCA is the best method to obtain the initial conditions. The advantage of unrotated PCA method is to replace the original complex variables with a few key linear combinations. However, despite that unrotated PCA can simplify the dimensions of complex variables; it may not be able to reasonably explain the physical phenomena. Unrotated PCA maximizes correlation coefficient sums of root mean square (RMS); whereas, Varimax [26,27] maximizes the variance of squared correlation coefficients in rotated principal components (RPCs). If the first-order moment is maximized, causing the correlation coefficients between unrotated principal components and measurable variables to form poor distribution and the correlation coefficients having no significant differences, the differences between unrotated principal components will thus be difficult to identify. This further makes it difficult to explain about the relation between variables and unrotated principal components, and interpret the physical meanings of unrotated principal components; whereas, if the second-order moment is maximized, the correlation coefficients between RPCs and measurable variables will then form wide distribution. Therefore, among any of the RPCs, few measurable variables have high factor loadings with most of them being zero, which makes it easy to explain correlation between measurable variables and the RPCs.

The time series values of pollutant concentration could first be normalized (Eq. (1)).

$$z_{ik} = \frac{C_{ik} - \mu_i}{S_i} \tag{1}$$

where Z_{ik} represents the kth time series and score Z from monitoring station i; C_{ik} stands for the kth time series concentration value from monitoring station i; μ_i stands for the mean value of concentration from monitoring station i; and S_i represents the standard deviation of monitoring station i. Eq. (2) demonstrates the relation between the RPCs and score Z:

$$z_{ik} = \sum_{j=1}^{n} L_{ij} P_{jk} \tag{2}$$

where L_{ij} represents the factor loadings of the jth RPCs from monitoring station i; and P_{jk} represents the kth observation of the jth RPC.

2.2. Cluster analysis

CA is one of the MSA methods. Its purpose is to classify the variables with the minimal variance into groups, namely, to divide similar variance or observed data into same clusters. The process of HCA is to: (1) define the similarity, (2) select the linking method and (3) determine the number of clusters. The difference between the variables can be quantified by defining the similarity. Before defining the similarity, the data must be standardized (based on the mean value and standard deviation of each monitoring station). Subsequently, the Euclidean distance between each cluster shall be calculated, assuming that there are n monitoring stations and the Euclidean distance will form a distance matrix, $n \times n$, as Eq. (3):

$$d_{ij} = \left[\sum_{k=1}^{n} \left| x_{ik} - x_{jk} \right|^2 \right]^{1/2} \quad (3)$$

After selecting the appropriate distance, the appropriate linking method must be chosen for the variable linking. The purpose of the link is to cluster all the monitoring stations, starting from linking two monitoring stations and ultimately having all monitoring stations form groups. After the distance matrix is formed, Ward's method is used for choosing the preferentially connected monitoring stations. The merging principle of Ward's method is that the RMS and acquisition from all monitoring stations first merge with the monitoring station with minimal variance. The following monitoring station will then be merged after the new cluster is formed. As the differences of mean values between the clusters may not be significant, the remaining question then becomes how to determine the number of major clusters. Stooksbury and Michaels [28] indicated that pesudo-F, pesudo-t^2, correlation coefficient and total root mean squared variation [29] can be useful methods to determine the number of clusters. Among which, pesudo-F is the ratio of the total cluster variation to the variation within the cluster; and pesudo-t^2 is the ratio of the RMS of the two clusters to the sum of the RMS of one cluster. Both correlation coefficient and the sum of total RMS use clusters with large gradients as the number of clusters. In the study, the sum of total RMS is used as the index to determine the number of clusters. The sum of total RMS is defined as Eq. (4):

$$\mathrm{TRMSD} = \sum_{l} \sqrt{(x_i - x_l)^2} \quad (4)$$

TRMSD represents the sum of total RMS; x_i represents the value of concentration from any monitoring station; x_l represents the mean value of the first cluster. For instance, if two monitoring stations are linked into one cluster, the mean value of the cluster will be the mean value of concentration of the two monitoring stations. Consequently, the total RMS will change as well. The number of clusters can be determined once the sum of total RMS and number of clusters in the CA process are plotted into graphs.

3. Results and discussion

3.1. Geographical zoning of water quality control zones

The geographical zoning of water quality control areas provides the competent authority with consistent management of the areas with similar characteristics of polluted sources and concentration trends of distinct pollutants. Among the water quality monitoring data of the study area, TP, coliform, SS and BOD have surpassed the standards of water quality management in this area. Therefore, this research sequentially takes the four WQPs and follows the sequence of monitoring stations as well as the date of monitoring, applying PCA and CA methods to conduct various data analyses. The objective is to build references for the spatial delineation of the water quality control areas.

3.1.1. Results of principal component analysis

From the water quality monitoring item, coliform, in the results of PCA (Table 1), the eigenvalue of the first seven RPCs of coliform are found higher than 1 (dominant principal components) with explained TC concentration variation of 76.6%. The factor loadings between RPC and physical parameters are favorable in further identifying the characteristics of different RPCs. Fig. 2 demonstrates the distribution of the factor loadings from different RPC of the WQP, coliform. The results of Fig. 2 indicate that the first RPC represent the six monitoring stations in the midstream of the Beishi River as the cluster is formed due to high factor loadings (explained variation of 16.5%); the second RPC principal component stands for Nanshi River and the monitoring

Table 1
Eigenvalues and explained variances of rotated components for four water quality indicators

Rotated components	TC		BOD		SS		TP	
	λ	σ (%)	λ	σ (%)	λ	σ (%)	λ	σ (%)
1	3.8	16.5	4.2	18.4	8.6	37.2	6.6	28.5
2	3.0	13.1	3.5	15.1	3.5	15.0	6.2	26.9
3	2.8	12.1	2.7	11.7	2.4	10.6	3.7	16.2
4	2.7	11.7	2.6	11.1	1.6	6.7	3.2	13.8
5	2.4	10.6	2.0	8.5	1.5	6.3	2.0	8.7
6	1.5	6.5	1.9	8.0	1.4	5.9	–	–
7	1.4	6.0	–	–	–	–	–	–
Sum (%)		76.6		72.8		81.7		94.0

λ: eigenvalues, σ: explained variances (%).

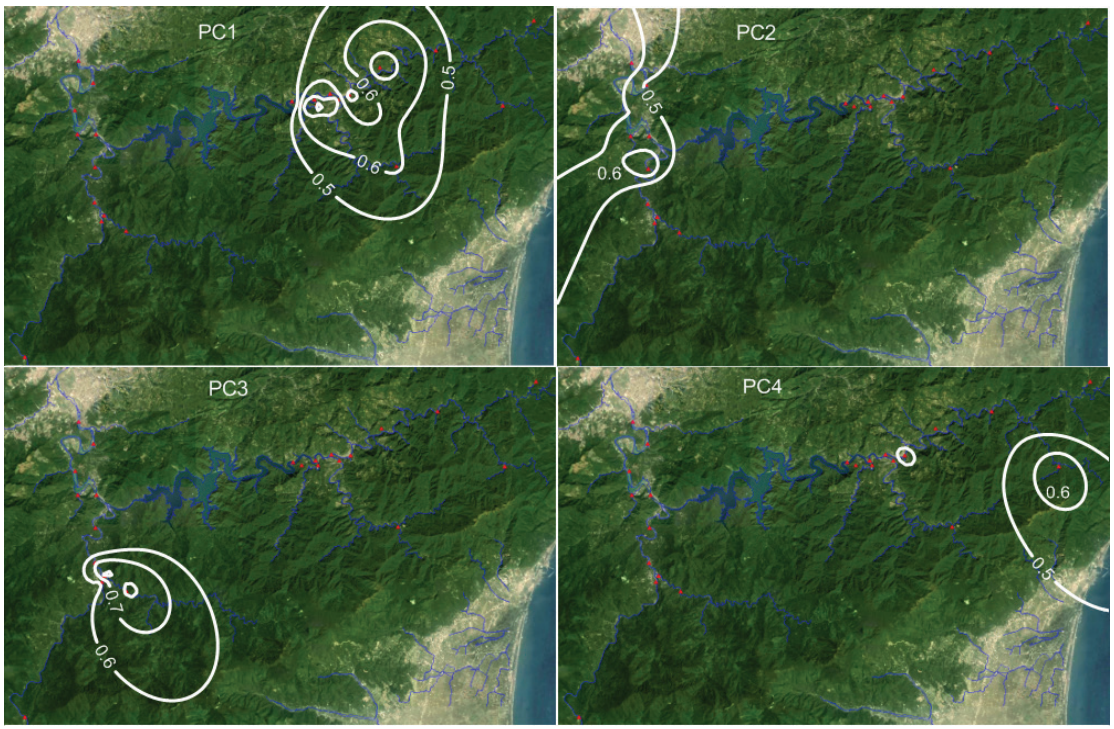

Fig. 2. Factor loading contours of rotated principal components for total coliform.

stations of Xindian River (explained variation of 13.1%), with the high factor loadings being concentrated at monitoring station B6; the third RPC represents the pollution formed in the hot spring area (explained variation of 12.1%), with high factor loadings being concentrated at monitoring stations B2, B3, B4, B5, etc., in the hot spring area; whereas, the fourth RPC represents sporadic monitoring stations with low pollution (explained variation of 11.7%).

Fig. 3 displays the intensity distribution of the factor loadings from various RPCs of WQP, BOD. The PCA results of BOD showed that the eigenvalues of the first six RPCs were higher than 1 with explained concentration variation of 72.8%. The first RPC represents the formation of group in the midstream of Beishi River (explained variation of 18.4%), with the high factor loadings being concentrated at monitoring station A2 and A5; the second RPC represents the monitoring station at Nanshi River (explained variation of 15.1%), with the high factor loadings being concentrated at monitoring station B2; the third RPC represents the pollution formed at the upstream of Xindian River and the downstream of Beishi River (explained variation of 11.7%), with high factor loadings being concentrated at monitoring station C2, C4, A6 and A12; the fourth RPC represents the low pollution of the sporadic monitoring stations (explained variation of 11.1%), with the high factor loadings being scattered at monitoring stations A10 and C1.

Fig. 4 shows the intensity distribution of the factor loadings from various RPCs of WQP, SS. The PCA results of SS indicated that the eigenvalues of the first six RPCs were higher than 1 with explained concentration variation of 81.7%. The first RPC represents high values at the Xindian River monitoring stations (explained variation of 37.2%); the second RPC represents the midstream of the Beishi River (explained variation of 15.0%); the third RPC represents sporadic pollution (explained variation of 10.6%), with the high factor loadings being concentrated at monitoring station A6 and A12; and the fourth RPC represents sporadic pollution (explained variation of 6.7%), with the high factor loadings being scattered at monitoring station A11 and C2.

Fig. 5 displays the intensity distribution of the factor loadings from PRCs of WQP, TP. The PCA results of TP indicated that the eigenvalues of the first five RPCs were higher than 1 with explained concentration variation of 94.0%. The first RPC representative is located at monitoring stations at Nanshi River and Xindian River (explained variation of 28.5%), with the high factor loadings being concentrated in monitoring station B6; the second RPC represents the sporadic pollution at the monitoring stations (explained variation of 26.9%), with the high factor loadings being concentrated in monitoring station A3; the third RPC represents the sporadic pollution at monitoring stations (explained concentration variation of 16.2%); and the fourth RPC represents the monitoring station at the midstream of Beidhi River (explained concentration variation of 13.8%), with the high factor loadings being concentrated in monitoring station A8 and A9.

Overall, the results of PCA specify that the trend of concentration variation of coliform, BOD and TP remains consistent; whereas, the PCA results of SS appear inconsistence with the three pollutants mentioned above. Moreover, the monitoring stations of Nanshi River (B2, B3, B4 and B5) are simultaneously divided as the third RPC of coliform, the second RPC of BOD and the first RPC of TP, with all high factor loadings. In addition, all of the four monitoring stations are located in the dense area of hot spring hotels (Fig. 6). This further explains that the distribution of hot spring hotels is related to the WQPs, TC, BOD and TP, and that the WQP, SS,

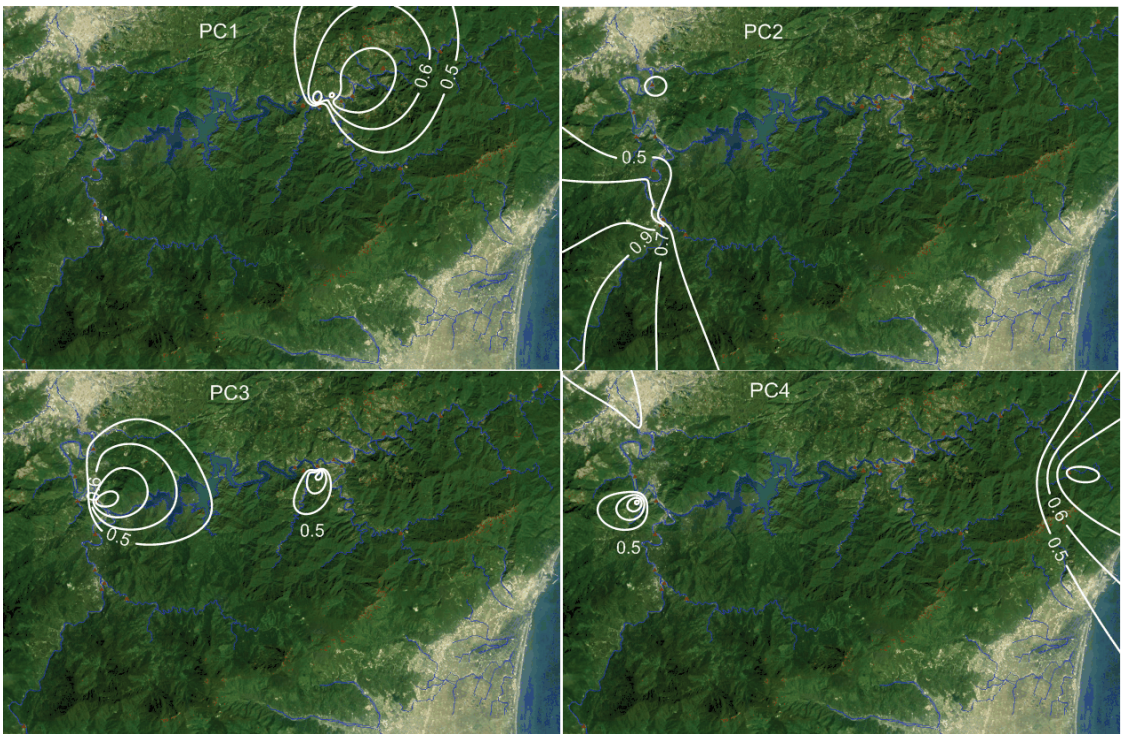

Fig. 3. Factor loading contours of rotated principal components for BOD.

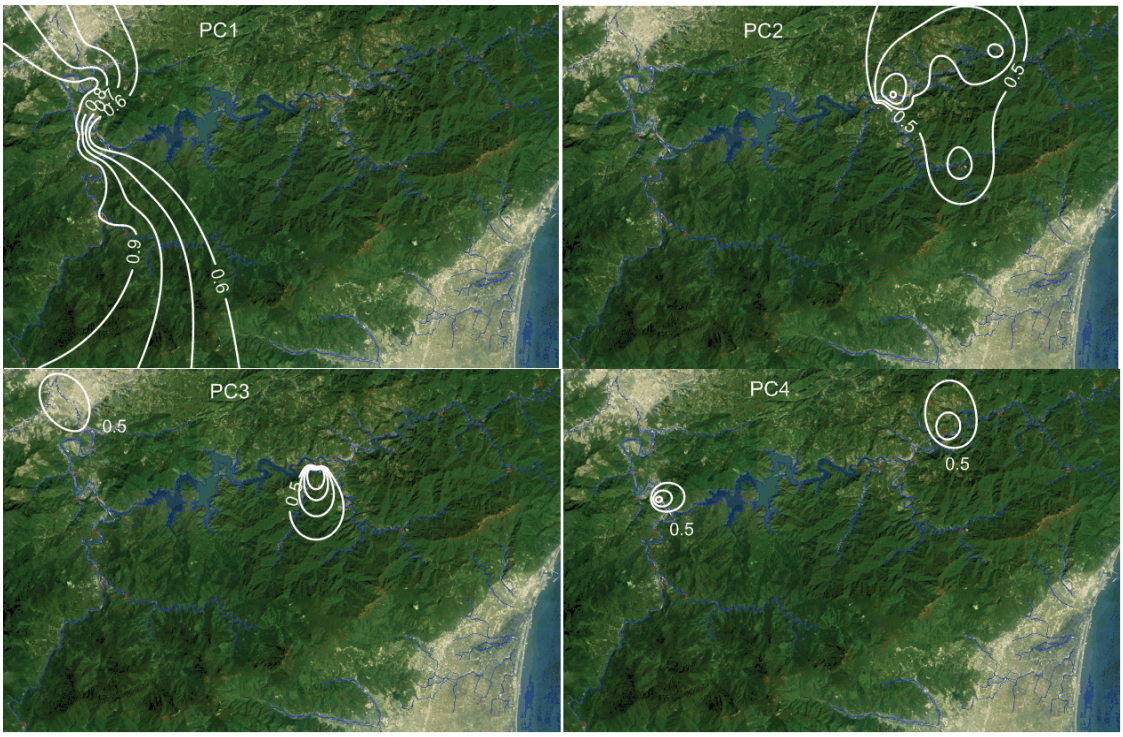

Fig. 4. Factor loading contours of rotated principal components for SS.

is mostly related to riverbank construction, rainfall or other artificial disturbances.

PCA method can effectively identify the characteristics and provide spatial delineation of water quality monitoring stations, namely that factor loadings of RPCs can pinpoint the group of monitoring stations with consistent concentration; however, PCA cannot be applied to divide all the monitoring stations into groups with low number of classification

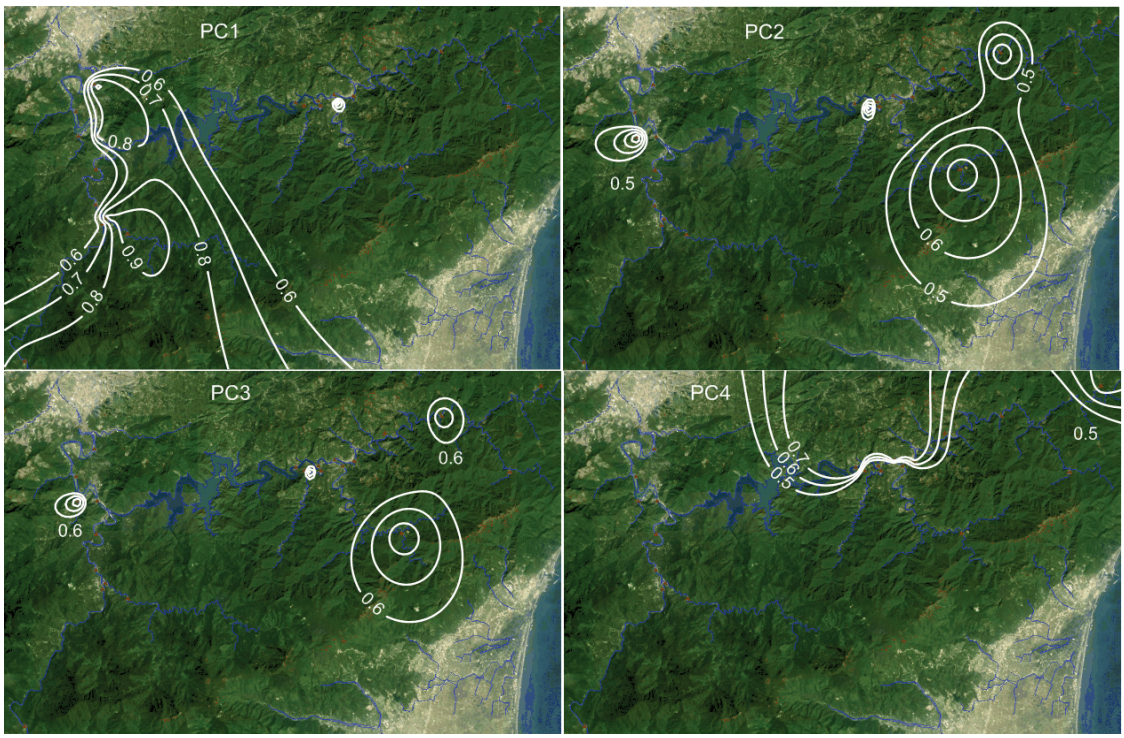

Fig. 5. Factor loading contours of rotated principal components for total phosphorus.

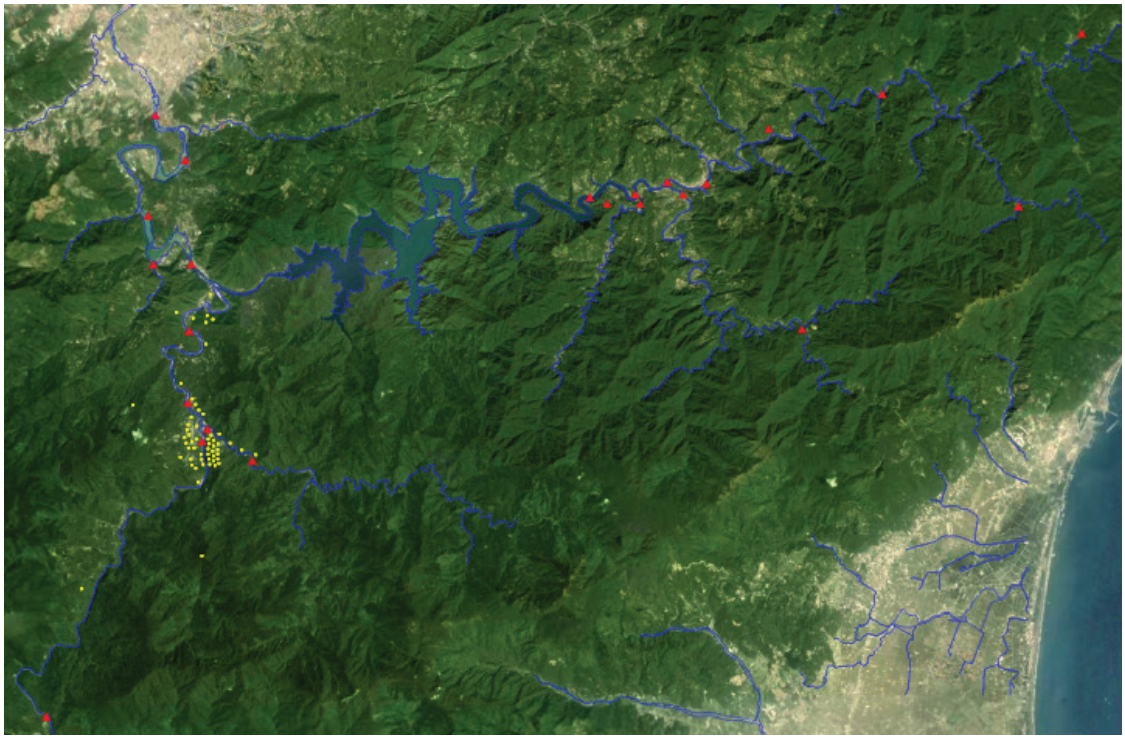

Fig. 6. Locations of hot spring inns and hotels (yellow dots) and monitoring stations (red dots).

when it comes to more complicated influential factors or pollutants with higher amount of key principal components. This makes it difficult to carry out spatial delineation for some monitoring stations in this research.

3.1.2. Cluster analysis

Several advantages of classifying monitoring stations by applying CA can be found: (1) providing an objective number of classifications; (2) dividing and classifying all

monitoring stations based on the correlation of WQPs over monitoring stations (correlation among the WQPs' concentrations of each station); and (3) dividing monitoring stations with lower consistency of WQPs' concentrations.

In fact, correlations of WQPs can be divided into several principal components by applying PCA. The more principal components are applied, the greater the variation can be explained; however, the number of principal components cannot be identified appropriately under this condition. According to the CA method applied in this research, the relation between the number of clusters and the explained variation of total RMS can be obtained (Fig. 7). Fig. 7 reveals

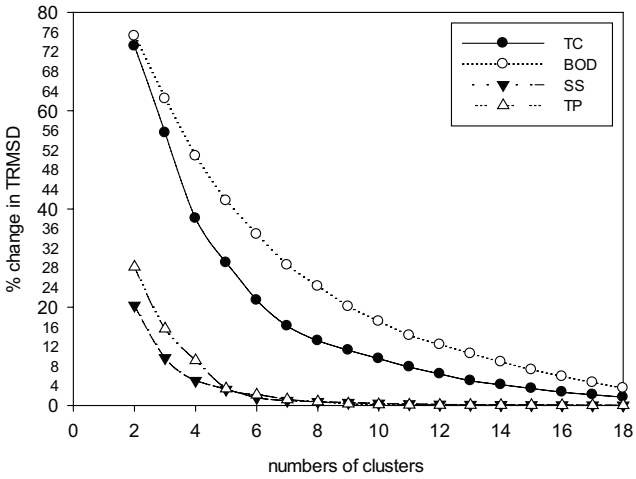

Fig. 7. Relationship between numbers of clusters and TRMSD.

that when the number of cluster increased from 4 to 5, the explained variation of total RMS of coliform reduced from 38.0% to 29.0%; BOD from 50.8% to 41.7%; SS from 9.1% to 3.4%; TP from 5.0% to 3.3%. For SS and TP, when the number of cluster increased from 4 to 5, the explained variation of total RMS can be increased less than 5%; therefore, the number of clusters suggested in this research is 4.

After analyzing the water quality data by applying CA, it was found that there are mainly two clusters of monitoring stations with consistent trends of WQPs' concentrations (small variation). The first cluster consists of the six monitoring stations at the midstream of Beishi River (A2, A4, A5, A6, A7 and A12). The trends of the concentration of coliform, BOD, SS and TP all appear to be consistent (Fig. 8). The second cluster includes the four monitoring stations at the downstream of the hot spring area (B2, B3, B4 and B5). Among which the concentration variation of the three WQPs, coliform, SS and TP, showed a consistent trend. As for the remaining monitoring stations, there is no common consistency observed in the concentration of the pollutants.

3.1.3. Comparison of spatial delineation by MSA

According to the results of PCA, the first two principal components of SS and TP explained the concentration variation to be 52% and 55%, respectively, which were the WQPs with lower degree of variation. However, the first two principal components of coliform and BOD explained 29% and 33% of concentration variation, respectively. By comparing the abovementioned WQPs via MSA, it can be discovered

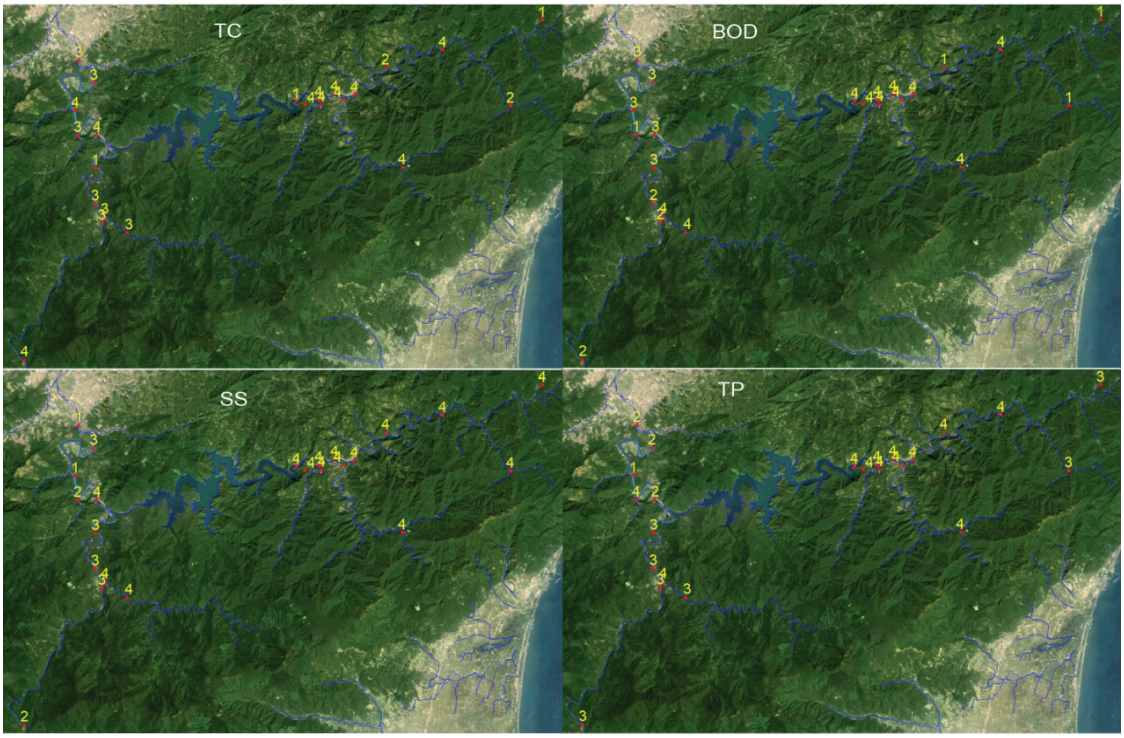

Fig. 8. Classification result for four water quality indicators with cluster analysis (the number over monitoring stations demonstrated the classification results, for example, 1 presented the first cluster and 2 presented the second cluster).

that the 23 stations are divided into four water quality control zones based on the trends of water quality variation, which are: (A) the midstream of Nanshi River (B2, B3, B4 and B5), where it is an intense area full of hot spring hotels. The main sources of this region include hot spring wastewater, domestic sewage and recreational activities; (B) the downstream of Nanshi River and Xindian River (monitoring station B6, C1, C2, C3, C4 and C5). For this area, there are many recreational water and fishing activities close to the city, and there are quite a number of recreational facilities along the river. The main sources are family sewage and recreational activities; (C) the midstream and downstream of Beishi River (monitoring station A1–A8, A11 and A12). The main sources of this region include recreational activities, wastewater from animal husbandry (deer farm) and agricultural wastewater (tea plantation); (D) the remaining sporadic monitoring stations (monitoring station A9, A10 and B1). This area consists of mainly ecological conservation area and primitive forest area. Only a few people are active and there is no particular source of pollution. Most of the WQPs of the monitoring station are far below the water quality standard.

According to the results of PCA, the midstream of Nanshi River (monitoring station B2, B3, B4 and B5) ranks to the third RPC for coliform, the second RPC for BOD and the first RPC for TP; whereas, the results of CA indicate that the area ranks to the third cluster for coliform and third cluster for TP. As the abovementioned four monitoring stations all are located in an area where hot spring hotels are densely situated, the results of PCA and CA specify that the wastewater from hot spring may affect WQPs, TC, BOD and TP. The main influential pollutant is TC, affecting areas where the abovementioned four monitoring stations at Nanshi River cover. The extent to which the WQPs are affected requires further investigation by conducting descriptive statistics, ANOVA test, PCA and CA.

3.2. Variation and characteristics of pollutant concentration in different water quality control zones

This research takes coliform as the research subject in order to understand the variation of respective pollutant concentration in the water quality control area. The analyses include descriptive statistics, ANOVA, PCA and CA. The results of CA and PCA have classified 23 monitoring stations into four water quality control zones, A, B, C and D. The mean values of WQPs of each water quality control zone and the results of ANOVA test (Table 2) reveal that there are significant differences among the 10 WQPs except for the temperature. In the downstream of Nanshi River and Xindian River basin (Cluster B), the highest WQPs are identified to be TC (2,532), BOD (0.96 mg/L), turbidity (16.22), NH_3–N (0.059 mg/L), TP (0.106 mg/L) and SS (10.88 mg/L). As for the monitoring station at the midstream of Nanshi River (Cluster A), all the mean values of WQPs were not significantly higher than that of the other three clusters. This signifies that the water quality did not deteriorate as it was not affected by the hot spring wastewater. Regarding the extent to which the monitoring stations in Cluster A were affected by the hot spring wastewater, it requires further ANOVA verification. In the midstream and downstream of Beishi River (Cluster C), the trend of WQIs did not remain consistent and the water quality of the sporadic monitoring stations (Cluster D) turned out to be the cleanest among the four areas.

Taking the concentrations of all pollutants in each cluster as samples and comparing the differences of four water control areas by PCA, it can be observed that the results of different clusters imply that the correlation between TC and other pollutants appears to be inconsistent (Table 3). The factors may be due to the differences in the sources of pollution in the four water quality control areas. The results of non-cluster showed the following explained variation for each parameter: SS and turbidity (19.6%), TC and NH_3–N (16.7%), BOD and COD (16%), and DO and temperature (16%). Therefore, the non-clustered TC and NH_3–N have a consistent concentration trend; whereas, the results from the monitoring stations at midstream of Nanshi River (Cluster A) displayed the following explained variation for each parameter: SS and turbidity (20.7%), DO and temperature (16.8%), BOD and COD (15.7%), NH_3–N and pH (13.8%) and TC (11.7 %). The TC of the midstream of Nanshi River (Cluster A) is not consistent with other WQIs. The results from the monitoring stations at downstream Nanshi River and Xindian River (Cluster B) showed that SS and turbidity (19.7%), TC and NH_3–N (18.0%), BOD and COD (16.3%),

Table 2
Mean values of water quality parameters for four water quality control zones

Item/categories	No.	*p* Value	Cluster A	Cluster B	Cluster C	Cluster D
Count	4,521		565	1,883	209	1,864
Percentage (%)	100.0		17.2	36.9	4.6	41.2
TC, NIEA E202.54B	2,516	0.000	1,792	2,532	1,760	1,067
DO, NIEA W422.52B	8.14	0.000	8.38	8.29	7.62	8.35
BOD, NIEA 510.55B	0.77	0.000	0.70	0.96	0.94	0.74
COD, NIEA 515.54A	3.15	0.000	2.83	3.36	3.83	2.91
pH, NIEA W424.52A	7.66	0.000	7.67	7.51	7.52	7.58
Temperature, NIEA W217.51A	21.77	0.099	21.24	22.15	22.74	22.10
Turbidity, NIEA W219.52C	11.88	0.000	16.01	16.22	3.20	3.02
NH_3–N, NIEA W448.51B	0.031	0.000	0.027	0.059	0.037	0.032
TP, NIEA W427.53B	0.076	0.000	0.037	0.106	0.029	0.065
SS, NIEA W210.57A	7.72	0.000	7.49	10.88	3.93	2.91

Table 3
Eigenvalues and explained variances of rotated components for four water quality control zones

Clusters	Item	MV	FL	Eigenvalues	EV (%)
NO	SS	Turbidity	0.94	1.76	19.6
		SS	0.92		
	TC and NH_3–N	MPN	0.86	1.50	16.7
		NH_3–N	0.85		
	BOD and COD	BOD	0.81	1.44	16.0
		COD	0.85		
	DO and temperature	Temperature	0.80	1.44	16.0
		DO	–0.88		
	pH	pH	0.97	1.06	11.7
Cluster A	SS	Turbidity	0.94	1.86	20.7
		SS	0.93		
	DO and temperature	Temperature	0.87	1.52	16.8
		DO	–0.78		
	BOD and COD	BOD	0.78	1.37	15.7
		COD	0.81		
	NH_3–N and pH	pH	0.71	1.24	13.8
		NH_3–N	0.85		
	TC	MPN	0.94	1.05	11.7
Cluster B	SS	Turbidity	0.94	1.77	19.7
		SS	0.93		
	TC and NH_3–N	MPN	0.88	1.62	18.0
		NH_3–N	0.88		
	BOD and COD	BOD	0.83	1.47	16.3
		COD	0.83		
	DO and temperature	Temperature	0.79	1.49	15.9
		DO	–0.88		
	pH	pH	0.97	1.06	11.8
Cluster C	TC and BOD and COD	MPN	0.77	2.73	30.3
		BOD	0.86		
		COD	0.92		
	SS	Turbidity	0.88	1.63	18.1
		SS	0.65		
	Temperature and pH	Temperature	0.86	1.25	13.9
		pH	–0.67		
	DO	DO	–0.90	1.14	12.7
	NH_3–N	NH_3–N	0.95	1.12	12.4
Cluster D	SS	Turbidity	0.86	1.54	17.1
		SS	0.79		
	DO	DO	–0.86	1.51	16.8
		Temperature	0.82		
	BOD and COD	BOD	0.79	1.43	15.9
		COD	0.83		
	pH and NH_3–N	pH	0.80	1.18	13.1
		NH_3–N	0.63		
	TC	MPN	0.96	1.06	11.8

EV: explained variance (%), FL: factor loadings, and MV: measurable variables.

DO and temperature (15.9%) and pH (11.8%). TC and NH_3–N in the monitoring stations (Cluster B) have a consistent concentration trend.

The results from monitoring stations at midstream and downstream of Beishi River (Cluster C) displayed the following explained variation: TC, BOD and COD (30.3%), SS and

turbidity (18.1%), pH and temperature (13.9%), DO (12.7%) and NH_3–N (12.4%). TC, COD and BOD in the midstream and downstream of the Beishi River (Cluster C) have a consistent concentration trend. The results from sporadic monitoring stations (Cluster D) showed the explained variation as the following: SS and turbidity (17.1%), DO and temperature (16.8%), BOD and COD (15.9%), NH_3–N and pH (13.41%) and TC (11.8%). The TC of the sporadic monitoring stations (Cluster D) appears to be inconsistent with other WQPs. The abovementioned results of PCA indicate that there was no consistent concentration trend of TC and other WQPs in the midstream monitoring stations of Nanshi River (Cluster A) and sporadic monitoring stations (Cluster D). However, the concentration trend of TC concentration NH_3–N in the downstream of Nanshi River and the monitoring stations at Xindian River (Cluster B) appeared to be consistent. The variation trend of TC, COD, BOD WQIs in the midstream and downstream of the North Potential Creek (Cluster C) is consistent.

The monitoring stations at the midstream of Nanshi River have the consistent concentration variation trend. The samples can be divided effectively into three groups with different TC concentrations via CA; whereas, via ANOVA, it can be observed that only the concentration of TC and DO has significant differences (Table 4). The remaining WQIs did not have significant differences. Hence, for the monitoring stations at the midstream of Nanshi River (Cluster A), the cluster with highest TC concentration (Cluster AF) (TC mean 9,033, occupancy rate 17.0%) had the lowest DO mean (7.91 mg/L) and pH (7.64), and the highest mean temperature (22.1°C) and NH_3–N (0.036 mg/L). The area around Cluster A is a hot spring area. The local hot spring is a carbonate spring with weak alkaline. The water temperature of the source of Wulai spring is 74°C and the pH value is approximately 7.65. Thus, the wastewater from hot springs could possibly lead to the results of highest TC, lower DO and pH in the samples of Cluster AF; the Cluster AE with lowest TC concentration (mean 541, occupancy rate 58.8%) turned out to have the highest DO (8.25 mg/L) and the lowest COD mean (3.08 mg/L). The water quality of Cluster AE did not appear to be significantly affected by the hot spring wastewater. As for the TC intermediate concentration Cluster AG (mean 2,733, occupancy rate 23.2%), the water quality showed an inconsistent trend and did not appear to be affected by the hot spring wastewater. From the above analysis, the hot spring wastewater may lead to 17% of the monitoring stations at Cluster A to increase TC values and slightly decrease DO and pH values. The results imply that the hot spring wastewater in Wulai area will affect the water quality of the midstream of Nanshi River (Cluster A) with the most highly affected WQP being TC.

Table 4
Mean values of water quality parameters over distinct clusters for Cluster A stations

Item/categories	*p* Value	Cluster AE	Cluster AF	Cluster AG
Count	565	332	96	131
Percentage (%)	–	58.8	17.0	23.2
TC	0.000	541	9,033	2,733
DO	0.001	8.25	7.91	8.04
BOD	0.686	0.77	0.73	0.79
COD	0.620	3.08	3.19	3.29
pH	0.615	7.66	7.64	7.69
Temperature	0.655	21.83	22.10	21.41
Turbidity	0.242	12.85	2.12	16.37
NH_3–N	0.132	0.029	0.036	0.032
TP	0.592	0.037	0.031	0.042
SS	0.446	7.78	5.60	9.13

4. Conclusions and suggestions

This research applies multivariate analysis (MA) to analyze the spatiotemporal variation and characteristics of WQPs of water quality monitoring stations at river basins, and simultaneously analyze the division and WQPs of water quality control areas. The study has demonstrated that MA can effectively help to analyze the variations and the geographical features of WQPs at these water quality monitoring stations. The results point out that only single MA method cannot provide reference for the classification of water quality monitoring stations. It is necessary to combine CA and PCA in order to provide a better reference for classification. The PCA and CA manners can effectively classify the monitoring stations into four water quality control areas. The results of PCA from different water quality control areas indicated that the concentration of TC can have different correlations with different pollutants due to the characteristics of geographical and sources. The CA was used to classify the WQPs at the midstream of Nanshi River and it was found that only mean values of TC and DO had significant differences. For the water quality of this area, wastewater of hot springs could cause 17% (Cluster A) of the monitoring stations at midstream of Nanshi River to increase TC while decreasing DO and pH slightly. The WQP that was affected mostly by wastewater of hot springs was TC.

References

[1] T.Y. Yu, I.C. Chang, Spatiotemporal features of severe air pollution in northern Taiwan, Environ. Sci. Pollut. Res., 13 (2006) 263–275.
[2] H. Sakihama, M. Ishiki, A. Tokuyama, Chemical characteristics of precipitation in Okinawa Island, Japan, Atmos. Environ., 42 (2008) 2320–2335.
[3] J.L. Wang, C.H. Wang, C.H. Lai, C.C. Chang, Y. Liu, Y. Zhang, S. Liu, M. Shao, Characterization of ozone precursors in the Pearl River Delta by time series observation of non-methane hydrocarbons, Atmos. Environ., 42 (2008) 6233–6246.
[4] T.Y. Yu, L.F.W. Chang, Delineation of air quality basins utilizing multivariate statistical methods in Taiwan, Atmos. Environ., 35 (2001) 3155–3166.
[5] R. Yang, D. Cao, Q. Zhou, Y. Wang, G. Jiang, Distribution and temporal trends of butyltins monitored by molluscs along the Chinese Bohai coast from 2002 to 2005, Environ. Int., 34 (2008) 804–810.
[6] K.P. Singh, A. Malik, S. Sinh, Water quality assessment and apportionment of pollution sources of Gomti river (India) using multivariate statistical techniques—a case study, Anal. Chim. Acta, 538 (2005) 355–374.
[7] H. Yongming, D. Peixuan, C. Junji, E.S. Posmentier, Multivariate analysis of heavy metal contamination in urban dusts of Xi'an, Central China, Sci. Total Environ., 355 (2006) 176–186.

[8] Y. Ouyang, Evaluation of river water quality monitoring stations by principal component analysis, Water Res., 39 (2005) 2621–2635.
[9] D. Dominick, H. Juahir, M.T. Latif, S.M. Zain, A.Z. Aris, Spatial assessment of air quality patterns in Malaysia using multivariate analysis, Atmos. Environ., 60 (2012) 172–181.
[10] B.K. Eder, J.M. Davis, J.F. Monahan, Spatial and temporal analysis of the Palmer drought severity index over the southeastern United States, Int. J. Climatol., 7 (1987) 31–56.
[11] B.K. Eder, A principal component analysis of SO_4^{2-} precipitation concentrations over the eastern United States, Atmos. Environ., 23 (1989) 2739–2750.
[12] B.K. Eder, J.M. Davis, P. Bloomfield, A characterization of the spatiotemporal variability of non-urban ozone concentrations over the eastern United States, Atmos. Environ., 27A (1993) 2645–2668.
[13] S. Vardoulakis, P. Pavlos, Sources and factors affecting PM10 levels in two European cities: implications for local air quality management, Atmos. Environ., 42 (2008) 3949–3963.
[14] R. Koklu, B. Sengorur, B. Topal, Water quality assessment using multivariate statistical methods—a case study: Melen River System (Turkey), Water Resour. Manage., 24 (2010) 959–978.
[15] R.L. Olsen, R.W. Chappell, J.C. Loftis, Water quality sample collection, data treatment and results presentation for principal components analysis – literature review and Illinois River Watershed case study, Water Res., 46 (2012) 3110–3122.
[16] F. Zhou, G.H. Huang, H. Guo, W. Zhang, Z. Hao, Spatio-temporal patterns and source apportionment of coastal water pollution in eastern Hong Kong, Water Res., 41 (2007) 3429–3439.
[17] Y.H. Yang, F. Zhou, H.C. Guo, H. Sheng, H. Liu, X. Dao, C.J. He, Analysis of spatial and temporal water pollution patterns in Lake Dianchi using multivariate statistical methods, Environ. Monit. Assess., 170 (2010) 407–416.
[18] S. Shrestha, F. Kazama, Assessment of surface water quality using multivariate statistical techniques: a case study of the Fuji river basin, Japan, Environ. Modell. Software, 22 (2007) 464–475.
[19] B. Zhang, X. Song, Y. Zhang, D. Han, C. Tang, Y. Yu, Y. Ma, Hydrochemical characteristics and water quality assessment of surface water and groundwater in Songnen plain, Northeast China, Water Res., 46 (2012) 2737–2748.
[20] H. Juahir, S.M. Zain, M.K. Yusoff, T.I.T. Hanidza, A.S.M. Armi, M.E. Toriman, M. Mokhtar, Spatial water quality assessment of Langat River Basin (Malaysia) using environmetric techniques, Environ. Monit. Assess., 173 (2011) 625–641.
[21] H. Razmkhah, A. Abrishamchi, A. Torkian, Evaluation of spatial and temporal variation in water quality by pattern recognition techniques: a case study on Jajrood River (Tehran, Iran), J. Environ. Manage., 91 (2010) 852–860.
[22] A. Astel, S. Tsakovski, P. Barbieri, V. Simeonov, Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets, Water Res., 41 (2007) 4566–4578.
[23] Water Resource Agency, Taiwan Ministry of Economic Affairs, Taiwan, Water Quality Monitoring for Taipei Water Supply Protected Watershed, MOEAWRA1020171, 2013.
[24] H.H. Harman, Modern Factor Analysis, The University of Chicago Press, Chicago, 1967.
[25] D.F. Morrison, Multivariate Statistical Methods, 2nd ed., McGraw-Hill Book Co., New York, 1976.
[26] H.F. Kaiser, The varimax criterion for analytic rotation in factor analysis, Psychometrika, 23 (1958) 187–201.
[27] J.D. Horel, A rotated principal component analysis of the interannual variability of the northern hemisphere 500 mb height field, Mon. Weather Rev., 109 (1981) 2080–2092.
[28] D.E. Stooksbury, P.J. Michaels, Cluster analysis of southeastern U.S. climate stations, Theor. Appl. Climatol., 44 (1991) 143–150.
[29] E. Brankov, S.T. Rao, P.S. Porter, A trajectory-cluster-correlation methodology for examining the long-range transport of air pollutants, Atmos. Environ., 32 (1998) 1525–1534.