Desalination and Water Treatment
www.deswater.com
47 (2012) 182–197
September
1944-3994/1944-3986 © 2012 Desalination Publications. All rights reserved
doi: 10.1080/19443994.2012.696797
Taylor & Francis
Taylor & Francis Group

Performance assessment of water quality monitoring system and identification of pollution source using pattern recognition techniques: a case study of Chaohu Lake, China

Yunfeng Xu[a], Chunzi Ma[a], Shouliang Huo[b], Beidou Xi[b], Guangren Qian[a,*]

[a]*School of Environmental and Chemical Engineering, Shanghai University, No. 333 Nanchen Road, Shanghai 200072, P.R. China*
Tel. +86 021 66137758; Fax: +86 021 66137758; email: grqian@shu.edu.cn
[b]*Chinese Research Academy of Environmental Sciences, No. 8 Dayangfang, Beijing 100012, P.R. China*

Received 21 October 2011; Accepted 6 February 2012

ABSTRACT

The present study intends to evaluate the properties of two pattern recognition methods, principal component analysis (PCA) and cluster analysis (CA), for a better management of the water quality monitoring systems, and then to identify those lake areas with similar pollution behaviors and possible pollution sources. These methods were employed to analyze the four indexes: chlorophyll-a, secchi depth, total nitrogen, and total phosphorus, and were compared with each other. The results indicated that the classification outcomes by the PCA were consistent with those by the CA. Twelve monitoring sites were classified into 5, 7, or 8 groups based on their similarity characteristics of the pollution level. In addition, the pollution sources in the Chaohu Lake were mainly exogenous pollution, derived from the four rivers into the lake. These facts demonstrated that the PCA and CA methods had a great application potential for a better management of the water quality monitoring system, and the present paper provids a case study for many other lakes in China.

Keywords: Chaohu Lake; Principal component analysis; Cluster analysis; Water quality monitoring system

1. Introduction

Chaohu Lake (Anhui province), located in the Yangtze River Basin, is one of the five largest freshwater lakes in China. It serves a variety of functions, including flood control, water supply, irrigation, transportation, fishery [1], and tourism. Therefore, the lake plays an overwhelmingly significant role in the regional socioeconomic development. However, with population expansion and rapid economic development, this lake too has experienced a cultural eutrophication process with the excessive nitrogen and phosphorus inputs since the late 1970s [2], causing an accelerated growth of algae [3,4] and a reduction in the transparency [5]. Some methods have been adopted to control eutrophication in the lake since 1984 [6], albeit with limited success.

To control water eutrophication and improve the water quality, a water quality monitoring system composed of 12 normal observation points was established. These points are distributed along different areas of the Chaohu Lake, few of them perhaps

*Corresponding author.

displaying a similar behavior. The annual monitoring work is carried out the Environmental Monitoring Station of the Anhui Province. Thus, it is necessary to explore the water quality monitoring system performance and more important by applying the principal component analysis (PCA) and the cluster analysis (CA).

The PCA is a technique that reduces the original variables to minority underlying factors [7] (i.e. independent uncorrelated variables) or principal components (PCs), which account for as much of the total original variance as possible. Through the application of the PCA, the multicollinearity problem probably implied between the original variables could be avoided [8,9]. In addition, CA (such as hierarchical CA) is an unsupervised pattern recognition technique [10], designed to detect hidden "groups" or "clusters" in a set of objects, so that the members of each cluster behave similarly to each other and the groups are maximally separated [9].

The above mentioned two approaches have been combined to explore the primary information from the original data in the previous reports. Vega et al. assessed the seasonal and polluting effects on the quality of river water by utilizing the PCA and CA [10]. Singh et al. employed the two approaches to evaluate the temporal/spatial variations and interpret a large complex water-quality data set which was obtained during the monitoring period of the Gomti River in the Northern part of India [11]. Shrestha and Kazama used both the approaches to assess the surface water quality in the Fuji river basin, Japan [12]. All of them confirmed that these two methods could supplement each other and that a combination of the two approaches could corroborate the feasibility approach for analyzing and resolving environmental problems. Nevertheless, very few researches have has applied the PCA and CA to manage the water quality monitoring system by analyzing the water quality indexes.

The four monitoring indexes, including chlorophyll a (Chla), secchi depth (SD), total phosphorus (TP), and total nitrogen (TN), have been chosen to describe the characteristics of a water quality monitoring system in the Chaohu Lake. Phosphorus and nitrogen are the essential nutrients that are necessary for the growth of algae in lakes [5,13] and serve as the primary causal factors for lake eutrophication. Generally, TP and TN are the main monitoring indexes owing to their stable measurement characteristics. In addition, Chla and SD are important response variables. Chla is the major photosynthetic pigment of algae and macrophytes, and most often is employed as an estimator of the algal biomass. SD can easily provide a lot of information on the lake water quality and, together with TP, TN, and Chla, has been routinely used as a measure of the lake strophic status [14].

The objectives of the present paper were to estimate the properties of the PCA and CA for a better management of the water quality monitoring system, and then to identify those lake regions with similar pollution behaviors and possible pollution sources through analyzing the four indexes.

2. Materials and methods

2.1. Study area

The Chaohu Lake (N31°25′–31°43′, E117°17′–117°51′) (Fig. 1) is a shallow eutrophic lake, which is located in the Anhui Province, southeast China. It presents the features of high terrain in the west, low in the east, and flat in the middle and flows from west to east. It covers a surface area of 760 km^2, with an average depth of 3 m. Owing to being a shallow lake and strong winds (with an annual average wind speed of 4.1 m/s), no seasonal stratification of the water column is observed [15]. Fengle River, Hangbu River, and Nanfei River are the three main tributaries in the watershed, accounting for more than 60% volume of the runoff [16]. The only drainage stream-Yuxi River [17] is permitted a direct water exchange between the lake and the Yangtze River.

The water quality of the Chaohu Lake is seriously influenced by the excessive nutrients inputs, mainly from the industrial and municipal wastewater, manure discharges, agricultural drainage, and roadways runoff, and sediment resuspension. The first three causes are caused due to an external pollution and the last due to an internal pollution.

2.2. Monitoring points

For the purpose of controlling the lake eutrophication, improving, and better managing the water quality in the Chaohu Lake, 12 monitoring points were established according to sampling specification. Fig. 1 shows the distribution of the monitoring sites along the Chaohu Lake. Among these sampling sites, five sites are set near the mouth area of main tributaries, i.e. the inlets of Nanfei River, Shiwuli River, Paihe River, Xinhe River, and Zhaohe River, respectively. The detailed characteristics of water quality monitoring stations in Chaohu Lake are listed in Table 1.

The samples were taken from each site underwater 0.5m every month for over 8 years (from January 2000 to June 2008). The data were provided by the Chinese Research Academy of Environmental Sciences.

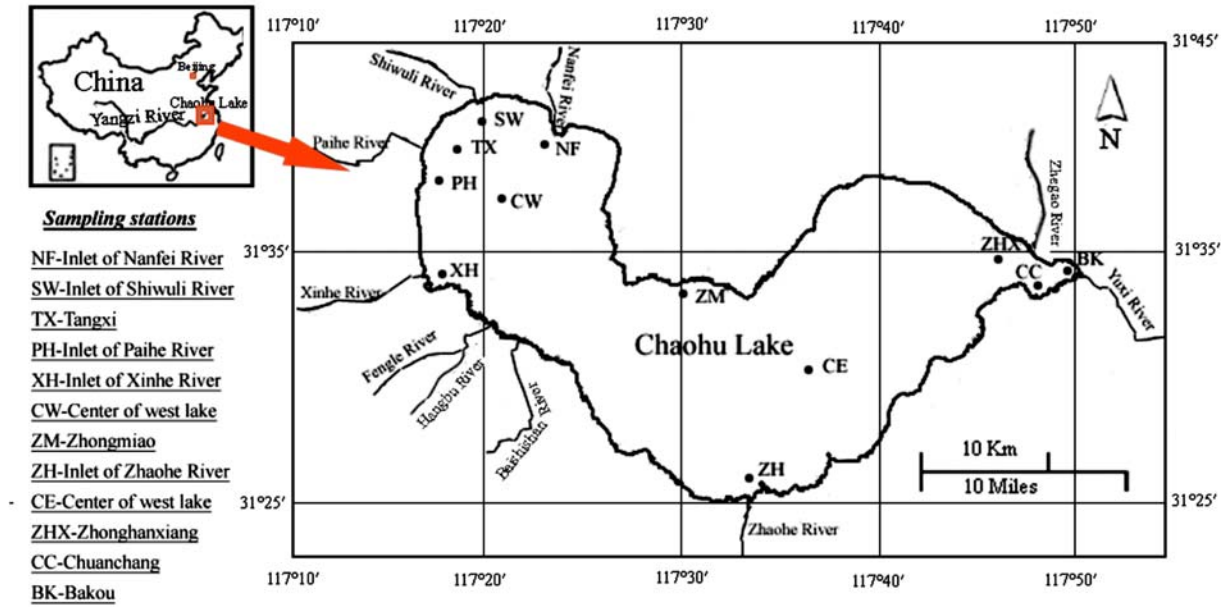

Fig. 1. A map of the study area and surface water quality monitoring sites in the Chaohu Lake.

2.3. Analytical methods

Ammonium Molybdate Spectrophotometry (GB11893-89, China) was employed to determine the TP Alkaline Persulfate Digestion Spectrophotometry (GB11894-89) which is used to measure the TN. Chla was extracted from the water body according to the acetone extracting agent and the colorimetric technique was used as a crude spectrophotometric method in the present paper. The SD was measured by utilizing the secchi disk and for the specific monitoring process one can refer to this document [19].

2.4. Pattern recognition methods

2.4.1. Principal component analysis

The PCA mainly creates new variables (so-called PCs) through a linear conversion. These PCs, which are linear combinations of the original variables, are orthogonal and uncorrelated to each other and keep maximum original information [9,20]. They are ordered in such a way that: the first PC accounts for the largest proportion of the original data variability, and then each subsequent one interprets the larger fraction of alterability that has not been explained by its predecessors. That is to say, most of the variation in the data set can be illustrated by the first few PCs [21]. These PCs are expressed by the following equation [22]:

$$PC_i = a_{1i}V_1 + a_{2i}V_2 + \cdots + a_{ni}V_n \quad (1)$$

where a is the component loading; V is the measured value of original variable, i is the component number, and n is the total number of variables.

To better illustrate the effect of each original variable in the PCs, varimax rotation is employed to obtain the rotated factor loadings that represent the contribution of each variable to a specific PC [23]. The varimax rotation ensures that each variable is maximally correlated with only one PC is minimally associated with the other components [21]. The rotated factor loadings of a variable are greater, the variable more contributes to the variation interpreted by the particular PC. In practice, only rotated factor loadings (with absolute values $\geqslant 0.5$ [24]) are chosen for the interpretation of the PC

2.4.2. Cluster analysis

The CA discovers an intrinsic structure or an underlying behavior of a data set without making any previous assumption about the data, to separate the objects into categories or clusters based on their similarity [10]. Hierarchical clustering is the most common approach in which the clusters are produced sequentially, through starting with the most similar pair of objects and forming higher clusters gradually. The Euclidean distance (defined by Eq. (2)) is usually used as a measure of similarity between the samples, and can be represented by the difference of analytical values from both the samples.

Table 1
Characteristics of the water quality monitoring stations in the Chaohu Lake

NO	Monitoring point	Detailed description
1	NF	Inlet lake district, industrial and municipal wastewater, manure discharges, agricultural drainage, and roadways runoff
2	SW	Inlet lake district, industrial and municipal wastewater, manure discharges, agricultural drainage, and roadways runoff
3	TX	Manure discharges, agricultural drainage, and roadways runoff
4	PH	Inlet lake district, industrial and municipal wastewater, manure discharges, agricultural drainage, and roadways runoff
5	XH	Inlet lake district, industrial and municipal wastewater, manure discharges, agricultural drainage, and roadways runoff
6	CW	Sediment resuspension
7	ZM	Manure discharges, agricultural drainage, and roadways runoff
8	ZH	Inlet lake district, industrial, and municipal wastewater, manure discharges, agricultural drainage, and roadways runoff
9	CE	Sediment resuspension
10	ZHX	Manure discharges, agricultural drainage, and roadways runoff
11	CC	Manure discharges, agricultural drainage, and roadways runoff
12	BK	Outlet lake district, municipal wastewater, and sediment resuspension

Note: Come from Qingying [16] and Chao and Qinguo [18].

$$d_{ij} = \sqrt{\sum_{k=1}^{m} \left(X_{ik} - X_{jk}\right)^2} \quad (2)$$

where X_{ik} is the measured value of the kth indicator of the ith sample; X_{jk} is the measured value of the kth indicator of the jth sample and d_{ij} expresses the Euclidean distance between the ith and jth samples. The smaller the value of d_{ij} is, the closer properties between the ith and jth samples are, and together will they group.

The number of clusters may be indicated graphically with a dendrogram, a tree diagram usually used in the hierarchical CA [25]. For a complete procedure of the clustering method used in the present study, we can refer to the literature [23].

3. Results and discussion

3.1. Principal component analysis

Although the PCA was mostly utilized to reduce the multiple dimensions, the method employed in the present paper was to serve as a non-parametric method of classification, so as to partition the monitoring sites into classes (PCs), which had a similar behavior and varied from those in other classes.

To understand the potential data structure, the number of PCs retained was identified in the Scree plot for Chla, SD, TN, and TP (seen in Fig. 2) [26]. It could be found that the first three eigenvalues expressed a greater slope, there was an unobvious change in slope after the third eigenvalue. In general, if eigenvalues were greater than 1 [27] and the PCs explained most of the variance in the original data set, these PCs were retained.

3.1.1. Chlorophyll a

The main results of the PCA application at all sites for Chla, SD, TN and TP are presented in Table 2.

From the table, it is evident that the first three PCs (with eigenvalues greater than 1) explained 86.86% of the information obtained from the original data set for analyzing the Chla mass concentration. This information displayed the data which were highly relevant and could be expressed by the first three PCs. Simultaneously, as the corresponding rotated factor loading values were <0.5, it has to be decided whether these sites could be clearly grouped into different classes in accordance with the variation of Chla. The first PC (PC1, 47.08% of variance) was heavily loaded by the contributions from BK, CC, ZHX, CE, ZM, and ZH sites, appearing to have similar characteristics in describing the level of Chla. In the PC2, the points of SW, TX and CW were combined to undertake 20.82% of the loading variance. PC3 (18.95% of variance) was mostly participated by PH, XH, and NF points.

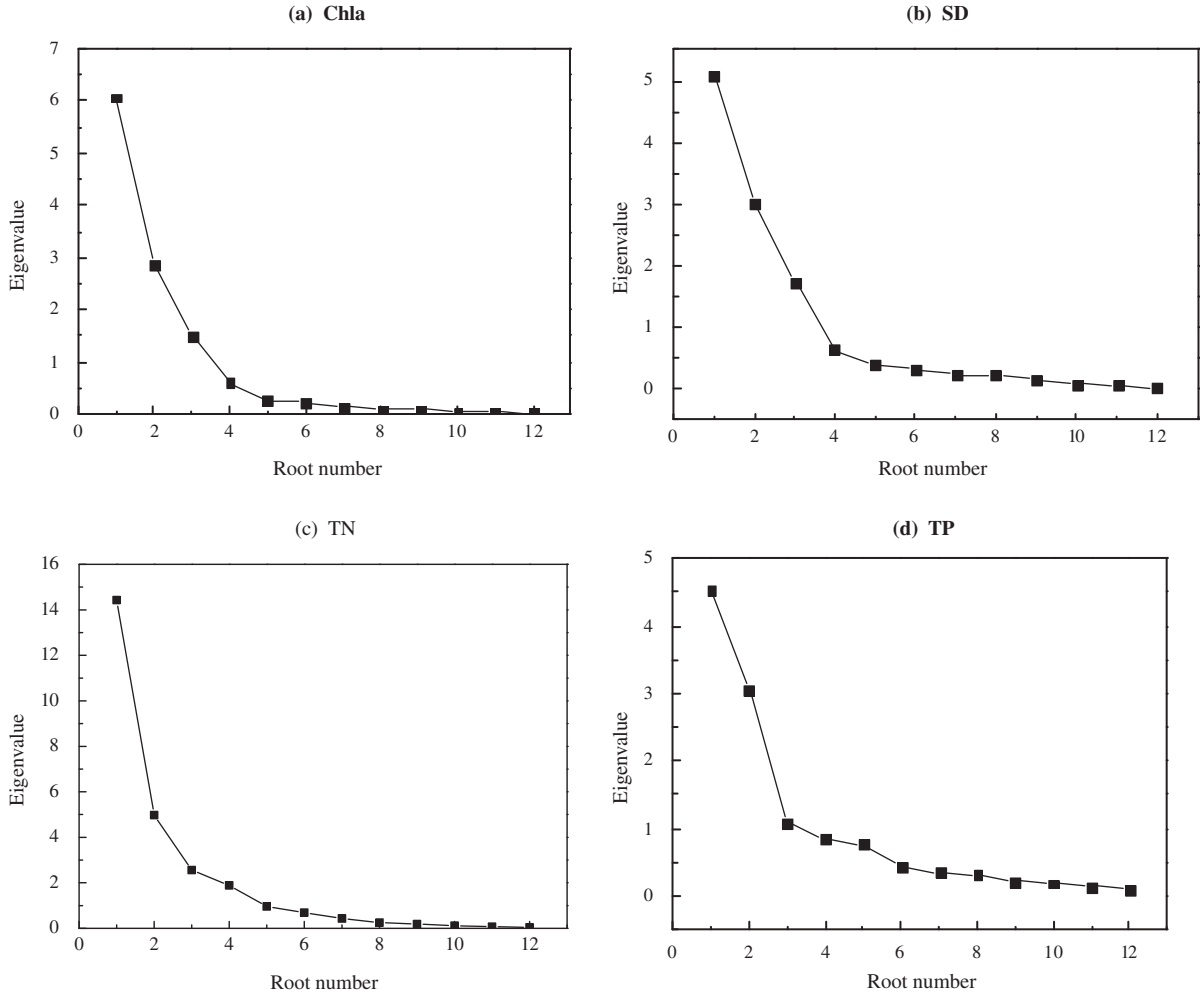

Fig. 2. Scree plot of the characteristic roots (eigenvalues) of PCs for Chla, SD, TN, and TP.

The loading values of the first three PCs were plotted in the load diagram, that of Chla is indicated in Fig. 3(a). It could be seen that similar results were achieved, i.e. the points associated with PC1, PC2, and PC3 got together automatically.

3.1.2. Secchi depth

The SD was a primary index for monitoring the water quality as well as Chla, so the PCA was applied to analyze the SD. Considering eigenvalues higher than 1, the first three PCs, which accounted for 81.98% of the data variance, were also required (seen in Table 2). PC1 had an important devotion of 42.43% from NF, SW, TX, PH, XH, and CW sites; while the loading variation of 25.02% was provided by PC2 whose contributions were from BK, CC, ZHX, and ZM sites. The CE and ZH points were significantly in accordance with PC3, its loading variation was 14.52%. The coincident results obtained from the load diagram are given in Fig. 3(b), where the points identified with PC1, PC2, and PC3 collected together automatically.

3.1.3. TN and TP

For TN and TP mass concentration, three PCs were selected for each index and their eigenvalues were all greater than 1, they explained 82.56 and 72.22% of the original variance, respectively. About TN, PC1 accounted for 54.24% of the variance and had an important dedication of SW, TX, PH, XH, and CW positions, while PC3 (9.61% of variance) was considered significantly in the NF sites. The other sites were mainly loaded by PC2 (18.70% of variance). The TX site was not only important for PC2, but also devoted to PC1, their factor loadings were 0.674 and 0.64. With regard to TP, PC1 made important devotions (35.05%

Table 2
The main results of the PCA application for Chla, SD, TN, and TP at all sites

Points	Chla			SD		
	PC1	PC2	PC3	PC1	PC2	PC3
NF	0.312	0.252	*0.690*	*0.835*	−0.003	0.131
SW	0.184	*0.828*	0.113	*0.935*	0.058	0.033
TX	0.055	*0.905*	0.143	*0.938*	0.025	0.034
PH	−0.007	−0.008	*0.957*	*0.971*	0.008	−0.031
XH	−0.019	0.314	*0.887*	*0.943*	−0.012	−0.034
CW	0.078	*0.879*	0.189	*0.896*	−0.016	−0.007
BK	*0.963*	0.086	0.055	0.039	*0.868*	0.226
CC	*0.960*	0.116	0.086	0.034	*0.850*	0.197
ZHX	*0.943*	0.090	0.068	−0.053	*0.876*	0.086
CE	*0.950*	0.111	0.057	0.040	0.328	*0.900*
ZM	*0.956*	0.066	0.070	0.020	*0.725*	0.295
ZH	*0.977*	0.110	0.055	0.029	0.348	*0.852*
Eigenvalue	6.068	2.853	1.502	5.092	3.003	1.743
variance (%)	47.085	20.822	18.952	42.433	25.022	14.523
Cumulative variance (%)	47.085	67.907	86.859	42.433	67.455	81.978
	TN			TP		
NF	0.351	0.156	*0.923*	*0.634*	−0.143	0.181
SW	*0.823*	−0.030	0.219	*0.783*	−0.106	0.069
TX	*0.674*	*0.640*	0.061	*0.907*	0.129	−0.064
PH	*0.905*	−0.064	−0.002	*0.898*	0.177	−0.076
XH	*0.831*	−0.014	0.082	*0.855*	0.124	0.07
CW	*0.778*	−0.009	0.142	*0.884*	0.101	0.121
BK	0.024	*0.701*	0.119	0.01	*0.898*	0.082
CC	0.016	*0.723*	0.108	0.159	*0.844*	0.248
ZHX	−0.024	*0.627*	−0.001	−0.033	0.483	*0.619*
CE	0.079	*0.667*	−0.041	0.065	0.253	*0.709*
ZM	−0.027	*0.510*	0.066	0.011	*0.722*	0.431
ZH	−0.075	*0.611*	−0.015	0.134	0.103	*0.895*
Eigenvalue	14.428	4.975	2.556	4.532	3.057	1.077
Variance (%)	54.244	18.704	9.611	35.052	20.445	16.721
Cumulative variance (%)	54.244	72.948	82.559	35.052	55.497	72.218

Note: Values in italic indicate the variables that mostly influence the correspondent PC.

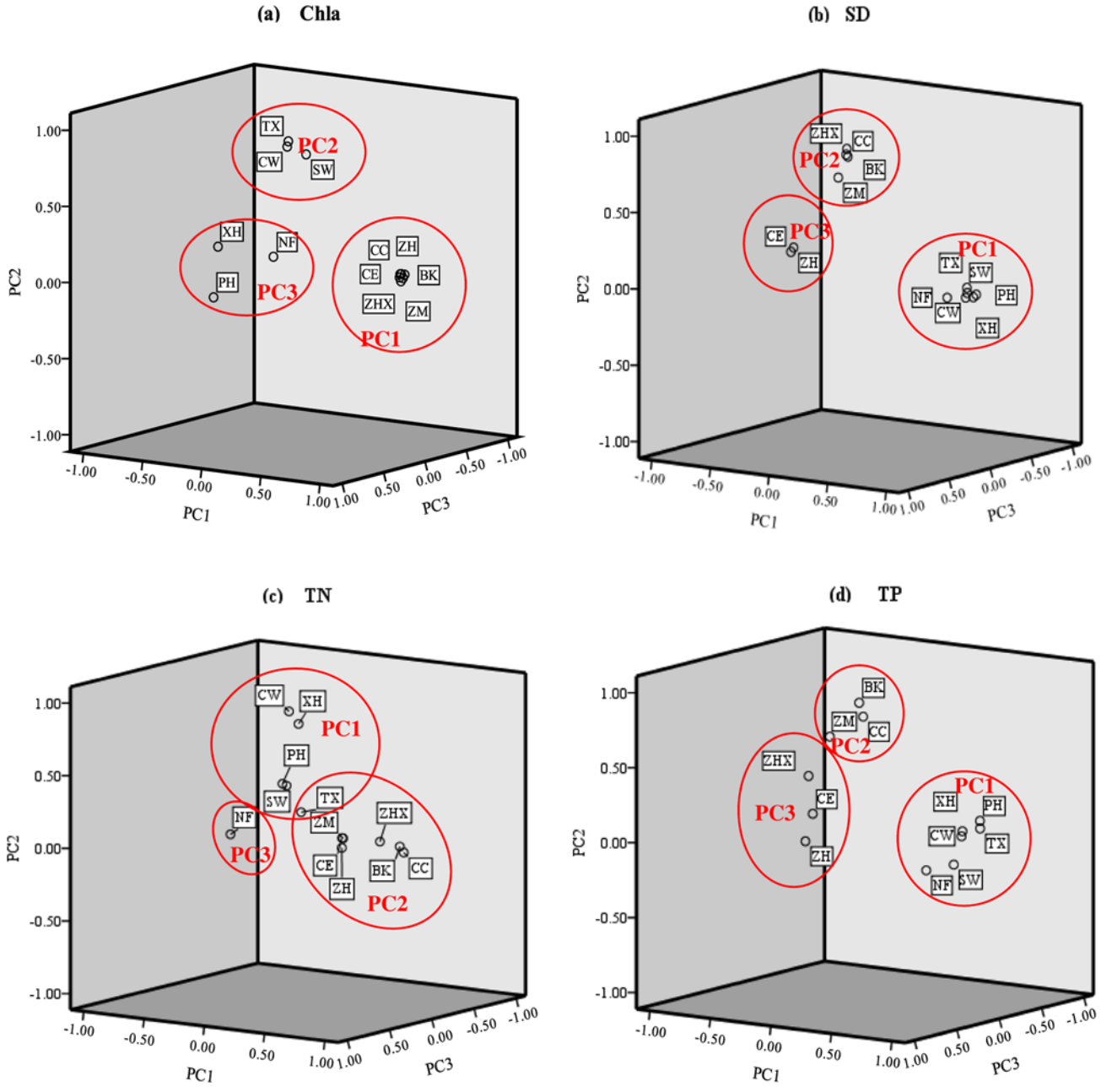

Fig. 3. PC plots of Chla, SD, TN, and TP in the rotational space (load diagram).

of variance) toward NF, SW, TX, PH, XH, and CW points, while in sites BK, CC, ZM, and sites ZHX, CE, ZH were evidently associated with PC2 and PC3, they contained 20.45 and 16.72% of the variance, individually. The results obtained from the load diagram in Fig. 3(c) and (d) agreed with PC1, PC2, and PC3.

Although these four indexes were all loaded with many sites and took up many percent for illustrating the variation, it was still difficult to distinctly classify them into different categories in detail to describe a similar behavior. Hence, it was necessary to apply the CA method to analyze these indicators.

3.2. Cluster analysis

In the CA method, the clustering procedure applied was the average linkage method. Euclidean distance was employed to calculate the distance between the monitoring points, and then the ratio of Euclidean distance was served as the Rescaled Dis-

tance Cluster Combine (RDCC). Fig. 4 expresses the dendrograms of CA for Chla, SD, TN, and TP.

3.2.1. Chlorophyll a

As can be seen from Fig. 4(a), the dendrogram of Chla indicated that 12 monitoring sites of the Chaohu Lake could be divided into 5 clusters: cluster 1 (A, BK, CC, ZHX, CE, ZM, and ZH sites), cluster 2 (B, TX, CW and SW sites), cluster 3 (C, NF site), cluster 4 (D, XH site), and cluster 5 (E, PH site). It could be illustrated that these results were consistent with those obtained from the PCA method, where the PC1 and PC2 were coincident with A and B, and PC3 corresponded to the C, D, and E clusters of the CA, respectively.

For the sake of showing the Chla's classification clearly, the relevant clusters in the points map are demonstrated in Fig. 5(a). The sites in cluster A were lay in the Central Eastern district of the lake, while the other points in B, C, D, and E, which were in the inlet of rivers, were all located in the West area. It could be concluded that the distribution of these clusters were evidently relied on the pollution level, which played a major part to produce this classification.

3.2.2. Secchi depth

According to the RDCC value of SD calculated (seen in Fig. 4(b)), 12 sites could be firstly grouped into three clusters: cluster 1 (A, sites NF, SW, TX, PH, XH, and CW), cluster 2 (B, sites BK and CC), and cluster 3 (C, sites CE, ZH, ZHX, and ZM). Similar results could be gained from the PCA method, where PC1, PC2, and PC3 corresponded exactly to A, B, and C, respectively.

When reducing the RDCC value, the C and A could be further sorted into different sub-clusters, such as C1 (BK and CC sites), C2 (ZHX site) and C3 (ZM site); A1 (XH and CW sites), A2 (TX, PH, and SW sites), and A3 (NF site). The distribution of these classified groups in the points map is presented in Fig. 5(b). It could also be discovered that the cluster of these sites was highly dependent on the pollution level as well as the Chla.

3.2.3. TN and TP

Likewise, the dendrogram from CA on the TN and TP mass concentrations and related clusters of point map are also shown in Figs. 4(c) and (d) and 5(c), and (d). It could be shown that the monitoring points were

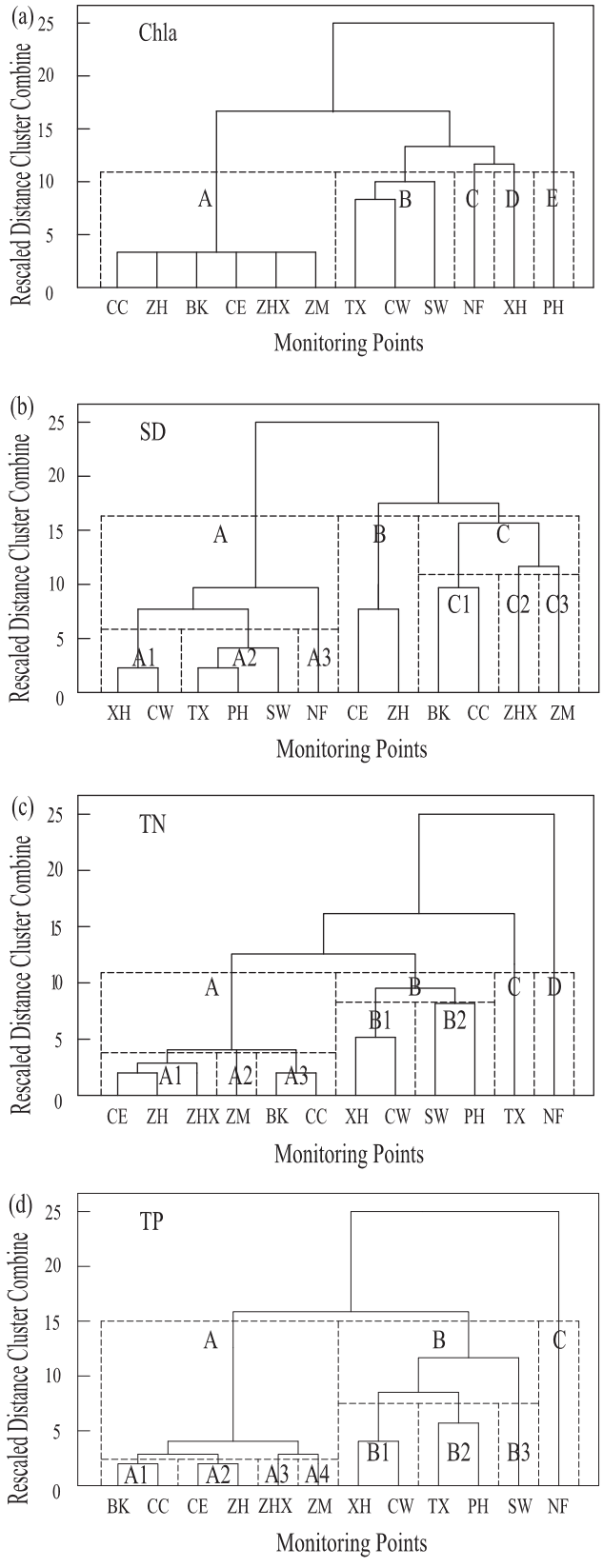

Fig. 4. Dendrograms for Chla, SD, TN, and TP.

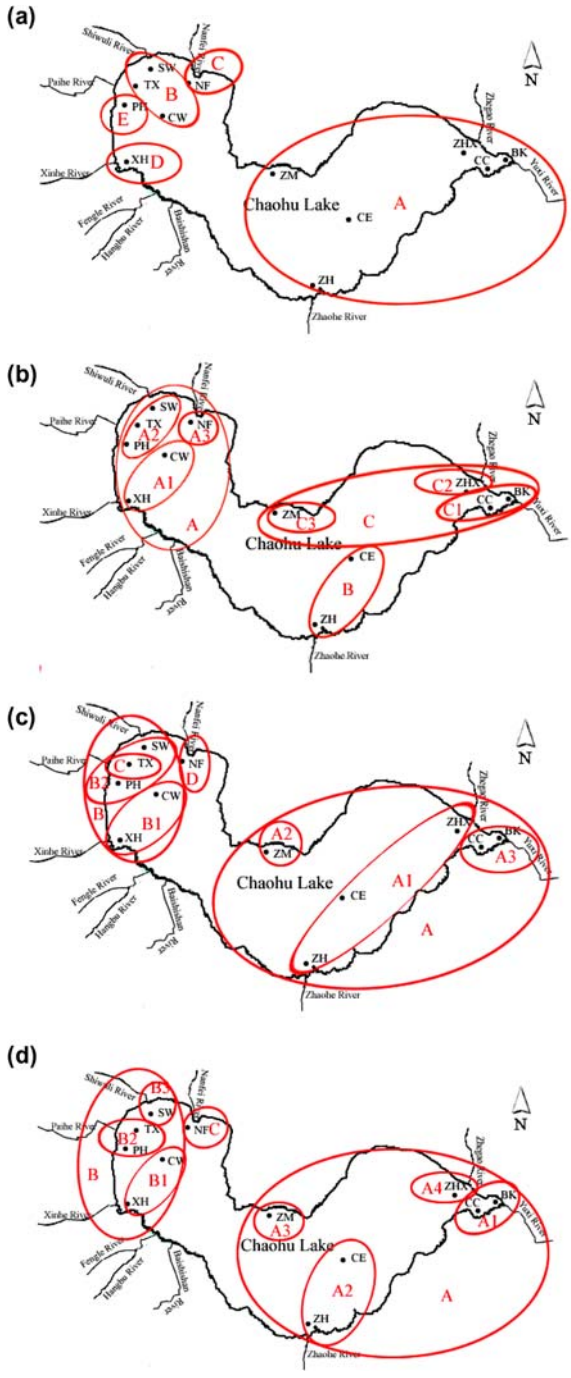

Fig. 5. Corresponding clustering point maps for Chla, SD, TN, and TP.

divided into different categories on the basis of the RDCC value.

For the TN concentrations, four categories, including A (CE, ZH, ZHX, ZM, BK, and CC sites), B (XH, CW, SW, and PH sites), C (TX site), and D (NF site), were firstly acquired from the CA method, which achieved similar results from the PCA, with PC1, PC2 and PC3 corresponding to A, B, and D, respectively. However, the site TX belonged to C category for its special properties, which had important contribution in both PC1 and PC2. In the same way, the A and B could be further classified into different sub-classes, containing A1 (XH and CW sites), A2 (TX, PH, and SW sites), A3 (NF site), and B1 (XH and CW sites), and B2 (SW and PH sites), through decreasing the value of RDCC.

Furthermore, for the TP mass concentrations, the 12 points were classified into three clusters: cluster 1 (A, sites BK, CC, CE, ZH, ZHX, and ZM), cluster 2 (B, sites XH, CW, TX, PH, and SW), and cluster 3 (C, site NF). From the results, it could be seen that the points gained in PC1 were divided into clusters B and C, and A included PC2 and PC3. To reduce the RDCC value, clusters A and B might be further divided into sub-clusters, i.e. A1 (points BK and CC), A2 (points CE and ZH), A3 (point ZHX), A4 (point ZM), and B1 (points XH and CW), B2 (points TX and PH), and B3 (point SW). These results declared that the PCA and CA achieved similar outcomes, that is to say, PC2, PC3, and PC1 were associated with A1 and A3, A2 and A4, B, and C, respectively.

According to Fig. 5, they demonstrated that there was a distinct description in the point map, and it could be concluded that the cluster of these sites was highly dependent on the pollution level as well as Chla and SD.

To prove the accuracy of the classification results from the CA method, the monthly average values of Chla, SD, TN, and TP at the categorical sites were analyzed in the next section.

At last, the classification results of PCA and CA for Chla, SD, TN, and TP at all sites are listed in Table 3. As can be found from the table, the classification results of the PCA were in agreement with those of CA, and compared with PCA, CA was a more specific sorting technique. Among these four indexes, the sites BK, CC, ZHX, CE, ZM, and ZH were remarkably separated from the other sites and they could be further classified into diverse sub-clusters. These sub-clusters may have been formed because the two types of points were quite different on the pollution level.

Via analyzing the CA outcomes, the points of BK and CC were always grouped into one category for all the indicators. In other words, these two points had similar characteristics of water quality indexes, they could be merged into one site. Meanwhile, the NF site was independent when comparing with the other points for all the indexes and was required to be monitored individually. The classification of the other points should be specifically discussed on the basis of the various water quality indexes.

Table 3
The classification results of PCA and CA for Chla, SD, TN and TP at all sites

Indexes	Methods	Category	Points											
			NF	SW	TX	PH	XH	CW	BK	CC	ZHX	CE	ZM	ZH
Chla	PCA	PC1							√	√	√	√	√	√
		PC2		√	√			√						
		PC3	√			√	√							
	CA	A							√	√	√	√	√	√
		B		√	√			√						
		C	√											
		D					√							
		E				√								
SD	PCA	PC1	√	√	√	√	√	√						
		PC2							√	√	√		√	
		PC3										√		√
	CA	A1					√	√						
		A2		√	√	√								
		A3	√											
		B										√		√
		C1							√	√				
		C2									√			
		C3											√	
TN	PCA	PC1		√	√	√	√	√						
		PC2			√				√	√	√	√	√	√
		PC3	√											
	CA	A1									√	√		√
		A2											√	
		A3							√	√				
		B1					√	√						
		C2		√		√								
		C			√									
		D	√											

(*Continued*)

Table 3 (*continued*)

Indexes	Methods	Category	Points											
			NF	SW	TX	PH	XH	CW	BK	CC	ZHX	CE	ZM	ZH
TP	PCA	PC1	√	√	√	√	√	√						
		PC2							√	√			√	
		PC3									√	√		√
	CA	A1							√	√				
		A2										√		√
		A3									√			
		A4											√	
		B1					√	√						
		B2			√	√								
		B3		√										
		C	√											

3.3. Data analysis

To prove the accuracy of the CA method, the monthly average values of Chla, SD, TN, and TP were analyzed in this section, in terms of the classification results of the CA. They are shown in Figs. 6–9, individually.

3.3.1. Chlorophyll a

From Fig. 6, the Chla averaged concentrations in cluster A ranged from 2.5 to 8.3 mg/m^3, while those in the B, C, D, and E clusters exceeded 16 mg/m^3, even could reach 103.8 mg/m^3. It appeared that the variation trends of the Chla averaged concentration were very similar in the A group, which agreed with the results gained from the PCA and CA methods. However, in the B group with a higher concentration, the sites a had the similar profile from 4 to 9 months and 10 to 12 months, while other months showed various changes.

Combined with Fig. 4, these results indicated that the similarity was highly dependent on the RDCC value. The RDCC values were lower within the category, the behaviors of the category were more similar. Toward clusters C, D, and E, the trends of monitoring sites were different, they seemed to have minimum similarity, and needed to be sample all alone.

3.3.2. Secchi depth

Simultaneously, on the basis of Fig. 7, the values of SD ranged between 30 and 47.3 cm in all points,

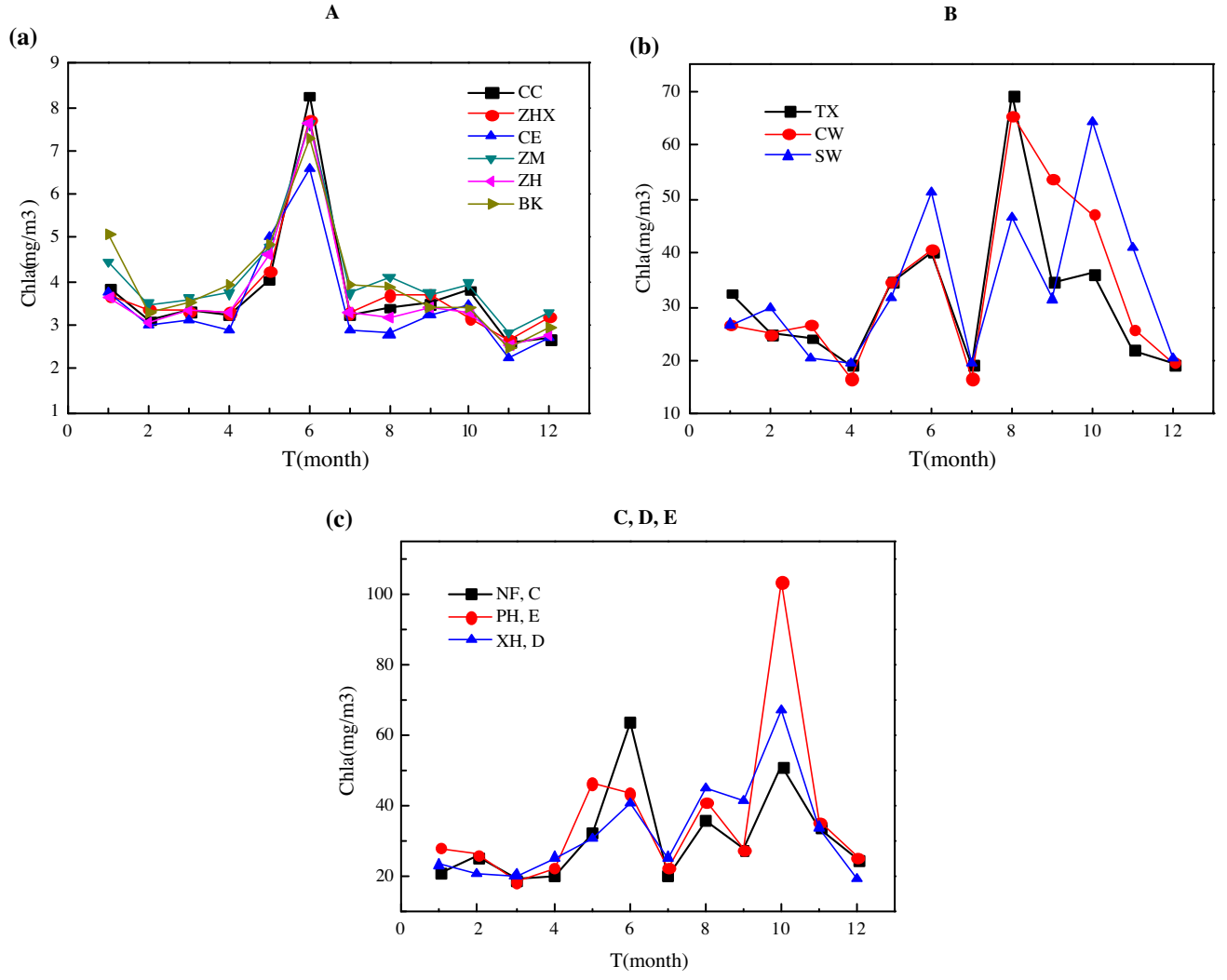

Fig. 6. The monthly average values from 2000 to 2008 for Chla at the monitoring sites, grouped by the correspondent CA category.

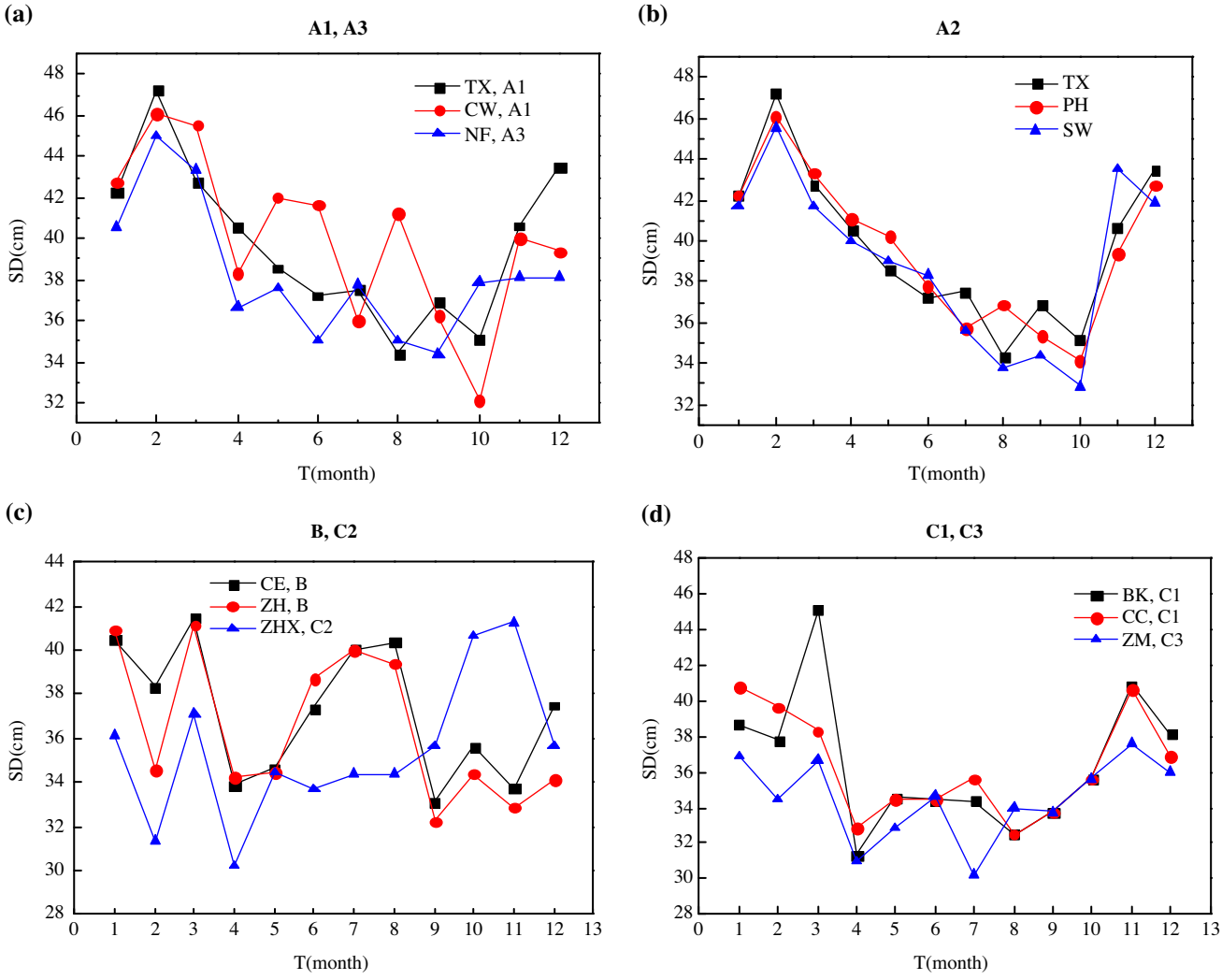

Fig. 7. The monthly average values from 2000 to 2008 for SD at the monitoring sites, grouped by the correspondent CA category.

and did not have distinct changes as well as Chla concentrations. This demonstrated that SD was not related to Chla directly. The high level of suspended sediment contented in Chaohu Lake was the main reason for the low SD [28]. For the A2, B, and C1 clusters, the monitoring sites had a similar profile. But for the A1 group, the sites did not have the same profile between 4 and 9 months, the other months had analogical features. With regard to cluster A3, C2, and C3, they seemed to have minimum similarity and should be monitored separately.

3.3.3. TN and TP

In addition, similar behaviors of TN and TP in sites could be observed in Figs. 8 and 9. The averaged concentrations of TN and TP in cluster A located in the east-central area were lower than those in other clusters, which were distributed along the western lake. These concentration variation trends were the same as those of Chla, which was because excessive TN and TP were beneficial to the growth of algae [6].

For TN, in cluster A1, the sites presented the same variation trends from 3 to 12 months, but the ZHX site showed a peak of TN in the first two months. This fact could be explained by the corresponding RDCC value of the ZHX site, which was greater than that of the other two sites belonging to the A1 category.

Similar conclusions could be drawn from Fig. 9. For TP, the sites in group A1, A2, and B1 all had a larger similarity intra-group, but the sites in the B2 group had a greater difference. These sites were related to

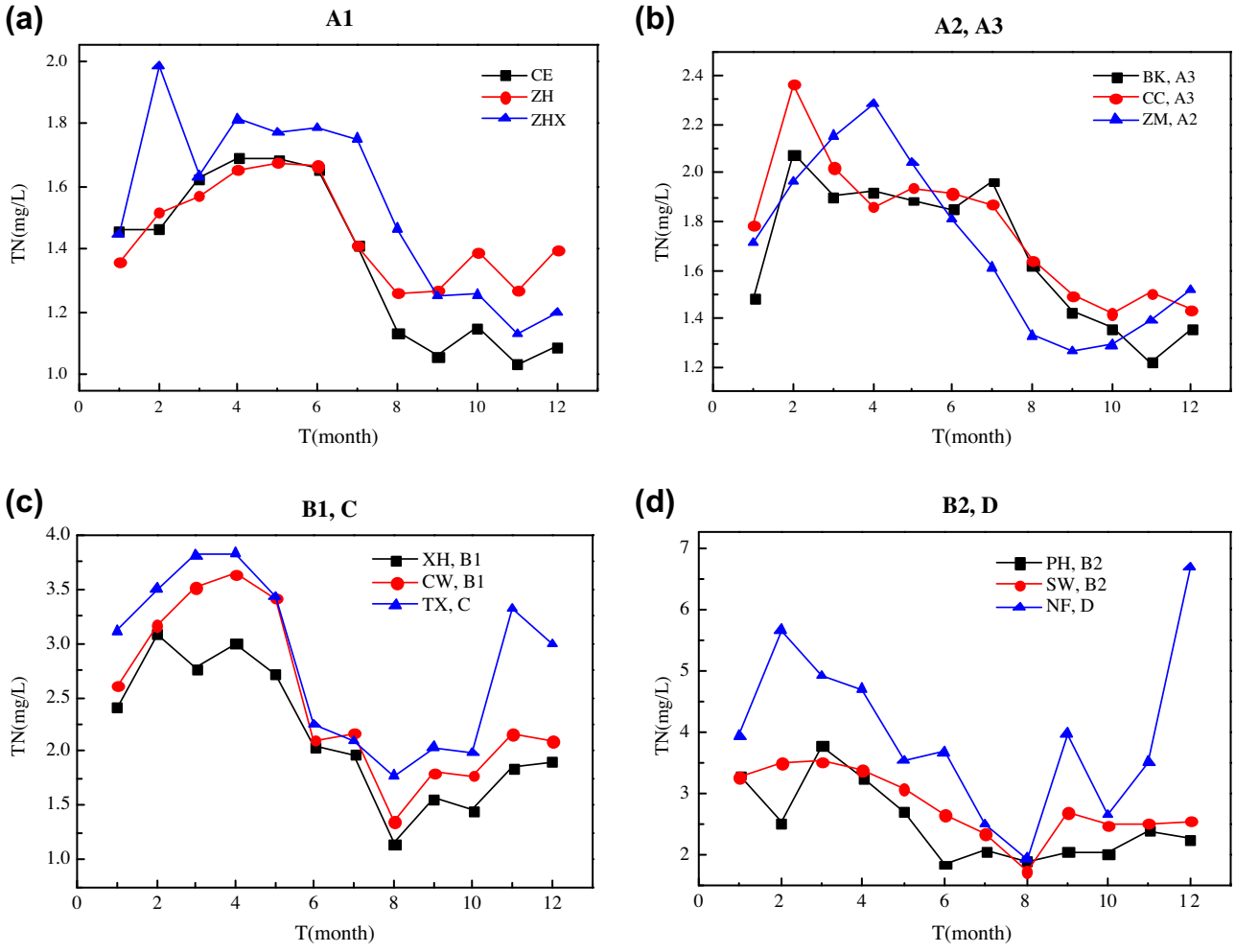

Fig. 8. The monthly average values from 2000 to 2008 for TN at the monitoring sites, grouped by the correspondent CA category.

their RDCC value, i.e. the value of B2 group was greater than that of A1, A2, and B1 group. The sites in group A3, A4, B3, and C appeared to have maximum distinction and should be classified individually.

In other words, these analyzing results confirmed that the Chaohu Lake had the same polluted behavior at many monitoring stations, this meant that only one monitoring point should be needed under the same pollution level. Moreover, they further verified that the western water quality of Chaohu Lake was in a much more deteriorated state than in the east-central area.

According to the data analysis and the discussion of water quality indicators—Chla, SD, TN, and TP by applying the PCA and CA methods, it was deduced that the sources of pollutants were mainly exogenous pollution, coming from the rivers into the lake. Endogenous pollution was not serious. The management of water quality in Chaohu Lake should control the entrance of external contaminant.

4. Conclusions

In the present paper, the PCA and CA methods were applied to the four water indexes at twelve sites in the Chaohu Lake. The conclusions obtained were as follows. (1) The identification of water monitoring sites by the PCA method was complied with that by the CA method. This indicated that PCA and CA methods had great application potential for better management of water quality monitoring systems. (2) The classification for these monitoring points evidently relied on the similarity of pollution behaviors, which played an important part. It was suggested that only one monitoring point should be needed to install under the same pollution level. (3)

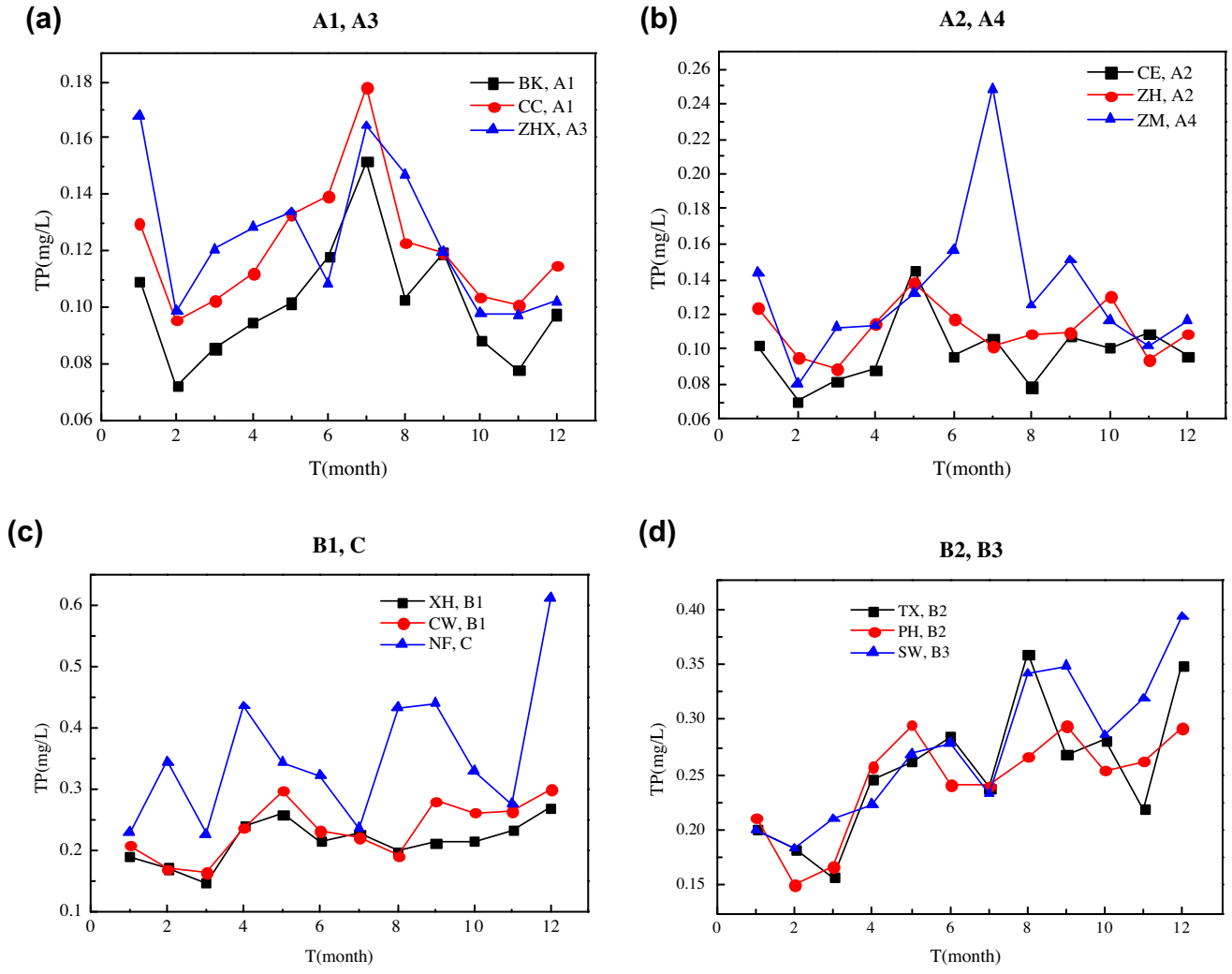

Fig. 9. The monthly average values from 2000 to 2008 for TP at the monitoring sites, grouped by the correspondent CA category.

It was inferred that the pollution source of Chaohu Lake was the exogenous pollution, derived from rivers into lake.

Therefore, this research would provide a case study to improve the management of water quality monitoring systems for many other lakes in China.

Acknowledgments

The work presented in this paper was financially supported by the Mega-projects of Science Research for Water Environment Improvement (No. 2009ZX07106-001).

References

[1] L. Guo, P. Xie, L. Ni, W. Hu, H. Li, The status of fishery recourses of Chaohu Lake and its response to entrophication, Acta Hydrobiol. Sin. 31(5) (2007) 700–705.

[2] S. Yao, S. Li, Sedimentary records of eutrophication for the last 100 years in Caohu Lake, Acta Sedimentol. Sin. 22(2) (2004) 343–347.

[3] R.D. Grundy, Strategies for control of man-made eutrophication, Environ. Sci. Technol. 5(12) (1971) 1184–1190.

[4] W. Rast, J.A. Thornton, Trends in eutrophication research and control, Hydrol. Process. 10(2) (1996) 295–313.

[5] R. Portielje, D.T. Van Der Molen, Relationships between eutrophication variables: From nutrient loading to transparency, Hydrobiologia 408–409 (1999) 375–387.

[6] G.-P. Shang, J.-C. Shang, Causes and control countermeasures of eutrophication in Chaohu lake, China, Chin. Geogr. Sci. 15 (4) (2005) 348–354.

[7] J. Huang, H.-D. Choi, P.K. Hopke, T.M. Holsen, Ambient mercury sources in Rochester, NY: Results from principle components analysis (PCA) of mercury monitoring network data, Environ. Sci. Technol. 44(22) (2010) 8441–8445.

[8] L. Wang, C. Wang, Z. Pan, Y. Sun, X. Zhu, Application of pyrolysis-gas chromatography and hierarchical cluster analysis to the discrimination of the Chinese traditional medicine *Dendrobium candidum* Wall. ex Lindl, J. Anal. Appl. Pyrol. 90 (1) (2011) 13–17.

[9] W.-Z. Lu, H.-D. He, L.-Y. Dong, Performance assessment of air quality monitoring networks using principal component analysis and cluster analysis, Build. Environ. 46(3) (2011) 577–583.
[10] M. Vega, R. Pardo, E. Barrado, L. Debán, Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis, Water Res. 32(12) (1998) 3581–3592.
[11] K.P. Singh, A. Malik, D. Mohan, S. Sinha, Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India)—a case study, Water Res. 38(18) (2004) 3980–3992.
[12] S. Shrestha, F. Kazama, Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan, Environ. Modell. Softw. 22(4) (2007) 464–475.
[13] G.F. Lee, W. Rast, R.A. Jones, Water report: Eutrophication of water bodies: Insights for an age old problem, Environ. Sci. Technol. 12(8) (1978) 900–908.
[14] G. Gibson, R. Carlson, J. Simpson, E. Smeltzer, Nutrient Criteria Technical Guidance Manual: Lakes and Reservoirs (EPA-822-B-00-001), United States Environment Protection Agency, Washington, DC, 2000.
[15] Q.Y. Tu, D.X. Gu, C.Q. Yin, Z.R. Xu, J.Z. Han, The Researches on the Eutrophication in Chaohu Lake, University of Science and Technology of China Press, Hefei, 1990.
[16] T. Qingying, G. Dingxi, Y. Chengqing, X. Zhuoran, H. Jiuzhi, A Series Researches on Lakes if China: The Chaohu Lake—Study on Eutrophication, University of Science and Technology of China Press, Hefei, 1990.
[17] X. Chen, X.D. Yang, X.H. Dong, Q.A. Liu, Nutrient dynamics linked to hydrological condition and anthropogenic nutrient loading in Chaohu Lake (southeast China), Hydrobiologia 661(1) (2011) 223–234.
[18] W. Chao, Z. Qinguo, Quantitative analysis of the impact factors on eutrophication of western Chaohu Lake, Master's thesis, Anhui Agricultural University, 2009.
[19] Water and Wastewater Monitoring and Analysis Methods, the Editorial Board of State Bureau of Environmental Protection, Water and Wastewater Monitoring Analysis Method, fourth ed., China Environmental Science Press, Beijing, 2002.
[20] C. Mendiguchía, C. Moreno, M.D. Galindo-Riaño, M. García-Vargas, Using chemometric tools to assess anthropogenic effects in river water: A case study: Guadalquivir River (Spain), Anal. Chim. Acta 515(1) (2004) 143–149.
[21] S.A. Abdul-Wahab, C.S. Bakheit, S.M. Al-Alawi, Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations, Environ. Modell. Softw. 20(10) (2005) 1263–1271.
[22] M. Statheropoulos, N. Vassiliadis, A. Pappa, Principal component and canonical correlation analysis for examining air pollution and meteorological data, Atmos. Environ. 32(6) (1998) 1087–1095.
[23] J.C.M. Pires, S.I.V. Sousa, M.C. Pereira, M.C.M. Alvim-Ferraz, F.G. Martins, Management of air quality monitoring using principal component and cluster analysis—part I: SO_2 and PM_{10}, Atmos. Environ. 42(6) (2008) 1249–1260.
[24] I.T. Jolliffe, Principal Component Analysis, Springer, New York, NY, 1986.
[25] J.E. Mckenna, An enhanced cluster analysis program with bootstrap significance testing for ecological community analysis, Environ. Modell. Softw. 18(3) (2003) 205–220.
[26] J.E. Jackson, A User's Guide to Principal components, Wiley-Interscience, John Wiley, New York, NY, 1991.
[27] S.M. Yidana, D. Ophori, B. Banoeng-Yakubo, A multivariate statistical analysis of surface water chemistry data—the Ankobra Basin, Ghana, J. Environ. Manage. 86(1) (2008) 80–87.
[28] W. Li, X. Wang, Y. Zhou, H. Wang, X. Li, Analysis of the suspended sediments spatial distribution and the reason in Chaohu Lake by use of TM images, Res. Soil Water Conserv. 02 (2006) 179–181.